A Statistics Primer

To

Advance Research Knowledge and Practice

In the Army Family Advocacy Program

A Collection of Articles from

Joining Forces/Joining Families Newsletter

1995–2007

Family Violence and Trauma Project
Center for the Study of Traumatic Stress
Department of Psychiatry
Uniformed Services University of the Health Sciences
Bethesda, Maryland
www.usuhs.mil/csts | www.centerforthestudyoftraumaticstress.org

Center for the Study of Traumatic Stress
Uniformed Services University of the Health Sciences
4301 Jones Bridge Road | Bethesda, MD 20814-4799
Tel: 301-295-2470 | Fax: 301-319-6965
www.usuhs.mil/csts | www.centerforthestudyoftraumaticstress.org

Design and production: Digital Design Group, Newton, MA, www.ddgdesign.com

Table of Contents

Topic *Page*

Introduction

This Statistics Primer is a compilation of articles on statistical methods that have been featured in the *Joining Forces Joining Families* newsletter. The newsletter, entering its thirteenth consecutive year of publication, is developed and disseminated by the Family Violence and Trauma Project (FVTP) of the Center for the Study of Traumatic Stress (CSTS). Importantly, Joining Forces Joining Families represents a commitment to fostering research within the Army Family Advocacy Program.

In 1995, FVTP collaborated with the Army Family and Morale, Welfare and Recreation Command (formerly known as the Community and Family Support Center) to bring knowledge to the FAP community of important research in the fields of family violence and child maltreatment, and to provide tools for understanding and conducting research. While FAP sites have unique characteristics in terms of population, geography and staffing, a common goal is to provide quality services informed by evidence-based practices.

Statistical methods are tools and a language by which research information is produced and shared. Research is the foundation for identifying and implementing evidence-based practices. Knowledge of statistics is fundamental for understanding the research outcomes of others and for conducting one's own research. The decision to use a procedure such as a measurement instrument or a clinical intervention requires a critical appraisal of whether it is likely to be successful or even helpful. Knowledge of measurement concepts and statistics will help in making decisions about program effectiveness.

"Building Bridges to Research" is the current name of the regular feature on statistical methods that appears in each issue of Joining Forces Joining Families. Whether you are a seasoned practitioner with longevity in FAP, or someone that is new to the Army's program and to clinical practice, we invite you to use the Statistics Primer to build your own bridges to research and to further your professional knowledge and work in the fields of family violence and child maltreatment.

We wish to recognize and thank the following people who have helped write and contribute to over a decade of *Joining Forces Joining Families* and its outreach: James E. McCarroll, PhD, Robert J. Ursano, MD, David M. Benedek, MD, Laurie E. Martin, ScD, John Newby, PhD, and Nancy Vineburgh, M.A.

Section 1

Basic Statistical Concepts

We start with a discussion of what statistics do: they describe or they estimate properties of distributions of numbers, which represent events. Central tendency and variability help to describe some of the characteristics of a distribution. When distributions have been described or estimated, they can be compared.

Central Tendency, Variability, and Comparisons

What kinds of statistics are there? One can broadly distinguish two types: descriptive and inferential (estimates). Among the tasks of statistics are to describe events and to determine whether results of such description can be generalized beyond the findings of the research performed.

Central Tendency

There are three main measures of central tendency: the mean, the median, and the mode. The mean is the arithmetic average; the median is the point above and below which half the observations fall; and the mode is the most commonly occurring observation.

Measures of Variability

Measures of variability tell you how much dispersion exists in a set of observations. These can be computed in many different ways. Two of the most common are the range and the standard deviation. The range is simply the distance between the highest and the lowest value in the set of observations. The standard deviation is an indicator of average variability from the mean. It is used in many other statistical measures (such as the t-test) as well as being descriptive in its own right. If you remember that there is less variability in a distribution with a small standard deviation than in one with a large standard deviation, you have the right idea. Another way to think of it is how close together the observations fall. Variability is lower if the observations cluster closely around one point than one in which they are more spread out. Variability is also important in determining statistical significance, a concept that will be later explored in more detail.

Comparisons

Comparison is the essential method of science. If you know how two groups of observations (distributions of observations) are described, you can make comparisons between measures of each. In statistics, questions are often framed in terms of the probability that an event happened by chance. Would this same finding be likely to happen if the study were repeated? When an investigator reports that a finding is "significant," this should mean that a statistical comparison has been made. Otherwise, some other word (such as "meaningful") is a better description of the result. A finding that is reported as significant usually means that the investigator has applied a statistical test that gives the probability that the same result would be likely to occur again if the test were repeated. If the probability is 95% *(confidence level)* that the same result would be obtained in x number of identical studies, the investigator would report that the findings are significant at the 0.05 level or above. This is a 5%-*significance level.* Similarly, a significance level of 0.01 (1%) indicates a confidence level of 99%.

Reference

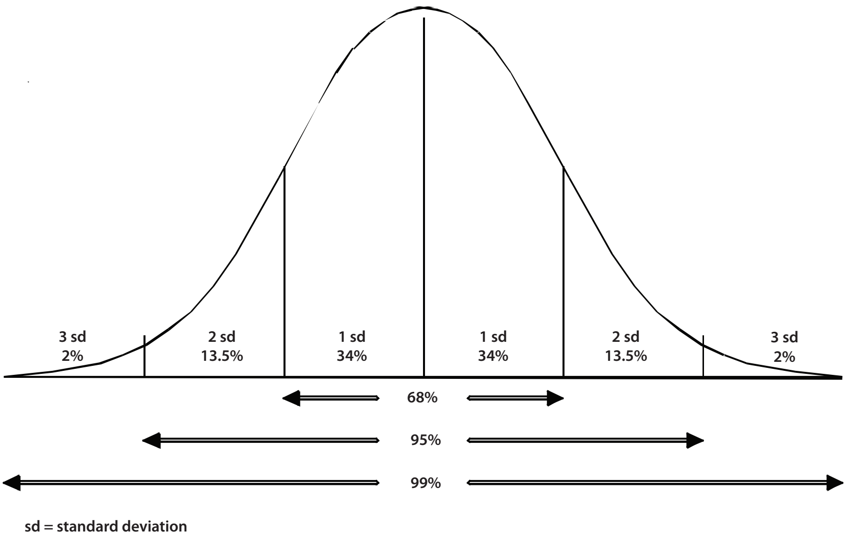
Fundamental Statistics in Psychology and Education. by J. P. Guilford Press, 1965.

Central Tendency: Mean, Variance, and Standard Deviation

When you measure certain phenomena, some measures cluster around a midpoint while others are more spread out. For example, take height. Height is usually thought to be normally distributed. A normal distribution means that the measures follow the distribution of a bell curve. It is wide at the bottom and rounded at the top. If you draw a line from the highest point at the top of the curve down to the bottom (the flat part of the graph), the point where they cross is called the mean. The rest of the distribution is spread out on either side of the mean. The extremes of each side are called the tails of the distribution because the area under the curve is very small and looks like a tail.

Many measures are distributed as a normal (bell) curve. For example, many women are about 5'5". As the heights get shorter or taller, approaching the two tails of the distribution, the percentage of women within each height category decreases. For example, there are fewer women who are less than

5' or taller than 6' than are 5'5". All measures are not normally distributed.



However, many sets of observations do have a normal distribution.

The variance and the standard deviation (sd) of a distribution of numbers are indexes of the spread of the measure in a sample or population. The standard deviation is the square root of the variance. To calculate a variance and a standard deviation, you need three numbers: the size of the group, the sum of each measure squared (the number multiplied by itself), and the square of the sum of all the measures. You can use the standard deviation to determine how much of your population falls within certain boundaries.

Using the mean and the standard deviation, it is easy to get a good description of the distribution of a measure in a population. In a normal distribution (a symmetric, bell shaped curve), 68% of the population is within one standard deviation of the mean (above it or below it). In other words, if you add the standard deviation to the mean one time, and subtract the standard deviation from the mean one time you will get the upper and lower boundaries for 68% of individuals in the population.

Suppose you wanted a range that includes a larger percentage of your population? Two standard deviations (i.e., the mean plus or minus the standard deviation multiplied by two) include 95% of your population. If you want to know the range that includes 99% of your population, you would again perform the same calculation except you would add to or subtract three standard deviations from the mean.

The mean and standard deviation are important statistical concepts that can help you understand the distribution of the sample you are working with and be used in other statistical tests such as the t-test.

For more information on these concepts, see; Koosis, Donald J. (1997). *Statistics: A Self-Teaching Guide.* John Wiley & Sons, 1997.

How Is Maltreatment Measured?

Descriptions of maltreatment are usually given in a numerical format: number of victims, how often victims are abused (e.g., so many every minute), number of homicides, and many others. Such descriptions may represent different concepts and types of measures. In this article we wish to draw a distinction between data that are obtained by (1) counting an existing population (e.g., the number of abused children annually reported to authorities) and (2) estimating from a sample (e.g., the number of spouse abuse victims, annually or lifetime, in the U.S.). Both can be referred to as statistics, but their sources and interpretation are very different.

Some events can be counted and expressed as frequencies (the number counted), proportions, percentages, or ratios. Examples of population count data are the annual report from the states on child maltreatment to the U.S. Department of Health and Human Services (published annually as *Child Maltreatment)* and the Federal Bureau of Investigation's *Uniform Crime Report* (http://www.fbi.gov/ucr/ucr.htm).

On the other hand, different types of statistics may provide a model of a phenomenon that is difficult or impossible to measure directly. There is no national spouse abuse reporting system (as there is for counting child maltreatment) and states vary in their laws, definitions, and mechanisms for reporting spouse abuse. Therefore, a population estimate seems to be a good way to describe the number of spouse abuse victims. Surveys have been conducted to estimate this number, but they are expensive, usually provide data on only one time point, and may suffer from methodological problems such as difficulty obtaining a representative sample. Two examples of population estimates are the Straus and Gelles (1986) and the Tjaden and Thoennes (2000) studies. Both were well-designed and well-conducted large-scale studies that provided population estimates of domestic violence using the Conflict Tactics Scale (CTS) (Straus, 1979), although different versions of the CTS were used in each study.

The person who wishes to compile and report statistical data on the frequency or rate of spouse or child maltreatment must pay attention to (1) the measure used, and (2) whether the frequency or rate is given for the sample studied or for the population as a whole.

References

Straus MA. (1979). Measuring intrafamily conflict and violence: The Conflict Tactics Scale. *Journal of Marriage and the Family, 41*:75–88.

Straus MA & Gelles RJ (1986). Societal change and change in family violence from 1975 to 1985 as revealed by two national surveys. *Journal of Marriage and the Family, 48*:465–479.

Tjaden P & Thoennes N. (November 2000). *Full report of the prevalence, incidence, and consequences of violence against women*. NCJ 183781.

U.S. Department of Health and Human Services, Administration on Children, Youth, and Families. *Child Maltreatment, 2003.* Washington, DC: U.S. Government Printing Office, 2005.

Section 2

The Chi-square Test and Statistical Significance

Following the description of distributions, variability, counts, and rates, we introduce the concepts of statistical significance and probability. The chi-square and the t-test are also introduced. We continue to explore the differences between counts and rates as ways of measuring maltreatment. Chi-square is one of the most basic statistical tests and it provides much information about data obtained from counting events. We give examples of the reasoning on which the chi-square test is based including how it is computed.

Significance vs. Meaningfulness in Statistics

Readers of research studies are often presented with the statement that a finding is statistically significant. What does this mean? Some people think that if a finding is statistically significant it is (1) true and (2) important. Neither of these is necessarily accurate. Statistical significance usually occurs in the context of a hypothesis-testing situation. When you perform a statistical test, you will find the value of a test statistic (e.g., the a value of the chi-square or t statistic or a correlation coefficient). The associated probability (p-value) tells you whether your hypothesis is supported by the data. When an investigator compares two or more groups and reports that a finding is statistically significant, it means that there is a certain probability (usually 95% or greater) that the finding did not occur by chance. (The language of probability is usually more precise than this, but we are simplifying it in this example.) Let's deal with the first of the two possibilities listed above, that a finding is true. As you can see, statistical significance is a statement about probability, not truth. Truth is not sought in statistics.

Now for the second issue, that of whether the finding is important. The sample size (as well as other factors) affects probability in statistics. The greater the number of subjects, the more likely you are to have significant findings. Thus, the findings may be significant, but not particularly meaningful because only a small difference is required to obtain significance with large samples. As pointed out by Lang, Rothman, and Cann (1998), a p-value does not convey unambiguous information because it is a mixture of confounded information: the size of the effect (related to your hypothesis), the size of the study (your number of subjects), and the precision of your measures. So, in addition to the p-value, you need to understand these other items: precision of the measure, number subjects, and the effect size. For example, suppose you have a good paper and pencil test (a precise measure) of some variable (such as depression or aggression scores) and your hypothesis is that there is a difference between men and women. Say that you find a value of 68.03 in a group of 15,000 men and 68.95 in a group of 15,000 women, and you have a p-value of less than 0.0001. Your result is statistically significant, but is it a meaningful difference? It may or may not be important, depending on the question you are asking. But, note that it is only a difference of 0.92. Suppose someone reported in a presentation of this study that there was a significant difference between men and women. At least two of your questions should be, "How much of a difference and what does that difference mean?" Answers to these questions would give you real information that you can use instead of just the knowledge that someone reported a significant difference.

Reference

Lang JM, Rothman KJ, & Cann CI. (1998). That confounded p-value. *Epidemiology, 9*:7–8.

Counts and Rates

The purpose of performing a study of an intervention or a prevention program is to determine if the program is effective. In order to do that, you have to select something to count. This may sound easy. However, there are other problems to consider. Here, the point is to illustrate a method to examine count data and its interpretation.

You can count a variable such as success or failure, or completed treatment versus non-completion of treatment or differences between men and women in the client population. If there are enough observations, you can perform some kind of statistical test. If counting is the only measurement possible, you are limited in the statistical tests you can perform. They may still be useful for your purposes and there is nothing wrong with them, but, if you can do more than count, you can perform more tests that will provide more information and may detect differences that more basic tests cannot.

For example, you can could the number of men and women who come to your class or your clinic and you can also measure them. You could give them a questionnaire or a test.

The chi-square test and the t-test are commonly seen in scientific publications and other reports. Each of these tests can tell you whether the hypothesis you have tested is statistically significant and the level of significance. When you can only count the variables or frequencies, chi-square is one way to test statistical significance. It is not the only test in this category and it has different variations. Here, it is used as an example of the kind of testing that can be done when you can only count events. If you can measure, other than by counting frequencies, you may apply more powerful tests, such as a t-test, to test the significance of the difference between the means of two groups.

In a study by Ethier, Lacharite, and Couture (1995), the authors compared the number of mothers scoring above the 90th percentile on a test of parental stress, Abidin's Parenting Stress Index (Abidin, 1983). They reported that there were significant differences (p<.0001) between negligent mothers and control mothers on one of the subtests, Child Domain. They counted the number of subjects (mothers) in each of four groups: *negligent mothers* whose scores were (1) above the 90th percentile, (2) below the 90th percentile, and *control mothers* whose scores were (1) above the 90th percentile and (2) below the 90th percentile. These counts were compared in a chi-square test and found to be statistically significant.

The authors also compared the mean score of negligent and control mothers on the Child Domain scale of Abidin's Parenting Stress Index. Since the measures were scores on a test and were more than count data, the authors chose to test the hypothesis that there was no difference between the negligent and control mothers using a t-test. They reported that the test was statistically significant (p<.001) and that they could reject the hypothesis of no difference with a high degree of confidence. Chi-square and t-tests were useful for the authors of the previously mentioned study. There are many tests of significance that can be described, but these are among the most common in social science. There are also extensions of the chi-square and t-tests test when you are testing the differences between more than two groups.

For example, analysis of variance is an extension of the t-test made applicable to more than two groups. This procedure allows for analysis of variables in a more efficient way than by performing a large number of comparisons between just two variables.

References

Abidin RR. (1983). *Parenting Stress Index.* Charlottesville, VA: Pediatric Psychology Press.

Ethier LS, Lacharite C, & Couture G. (1995). Childhood adversity, parental stress, and depression of negligent mothers. *Child Abuse & Neglect, 19*:6119–632.

Counts and Rates: Which Is More Important?

What is the question the commander often asks you? “How are we doing”? This is true whether the question is about FAP statistics, drug and alcohol use, or other indicators of “good order and discipline.” What is your best answer? It depends! If the inquiry were about FAP, would you answer relative to the counts (frequencies) or the rates (reference to a population) of maltreatment? It is often thought that the rate is the better answer because it includes the count and the population at-risk. Let us distinguish between a frequency (count) and a rate. The frequency is the count of cases. For example, the number of child abuse cases at Fort Installation during 2006 was 132. Of these, there were seven cases of child abuse with major physical injury.

A rate is based on two figures, a numerator and a denominator. The numerator is the number of cases. The denominator is the population at risk. In this case, “population at risk” is defined as the number of people capable of becoming a case (e.g., the population of children below age 18).

Suppose your commander wants details about cases that have been classified as physical abuse and you want to fully answer the question. What data do you need? First, you need the count of cases of physical abuse for the time period in question, say last year. Second, you need to know the size of the population at risk. Now that you have the frequency (number of cases) and the size of the population at risk, you can calculate a rate, say the rate per 1,000 children. Third, you need to know if there were any events that might have affected your numbers such as a change in the post environment, changes in the reporting rules or standards, new members on the CRC that may make their views known in definite ways or anything else that might have a bearing on your results.

Let’s plug in some numbers. In 2005 you found 6 cases of major physical injury to children in a population of 5,500 children. In 2006, there were 7 cases of major physical injury in a population of 5,000 children. So, you had a population decrease in the number of children and your number of cases as increased by one. Now, what are the rates and how do you calculate them? The rate per thousand is a simple proportion. If you have 6 cases for a population of 5,500 in 2005, what would the rate be if you had only 1,000 children? You divide 6 by 5,500 and then multiply by 1,000. Your rate per thousand is 1.09 in 2005 and 1.4 per thousand in 2006. In this case, is the rate the most meaningful statistic? It is probably misleading and you would be better off quoting the frequencies. Why? If you tell your commander you had this increase in the rate, it may not be very meaningful. The picture would probably be magnified beyond what you want to describe. I would report that the number of cases (frequency) was about the same. Find out the circumstances of the cases and see if there are any patterns. If, in 1997 your population of children decreases to 3,000 and you have 10 cases, something is happening and you need to determine what it is so you can apply a remedy. There is no rule for when to use a frequency and when to use a rate. Common sense may be the best guidance. Look for continuity in numbers (trends) and have some sense of the facts behind the numbers.

Understanding Rates Per 1,000 Using the Four-Fold Table

We continue the discussion on counts and rates considering what more they can tell us besides some quantity per 1,000. Knowledge of this procedure will help answer the commander's question, "How are we doing?" Suppose one has a population of persons in which it is known that either spouse or child abuse occurs. You can construct a table with four separate cells and four margins, called a four-fold table. Suppose that you want to compare the rate of child abuse among right-handed persons with that of left-handed persons for one year. You find in your cases that you have 75 right-handed persons and 15 left-handed persons. You also know that there are a total of 2,500 right-handed and 650 left-handed persons on post who were not involved in maltreatment cases. (These numbers are fictitious and were created solely for the purpose of this exercise.) You put this information into your table (see below).

	R-handed (Index)	L-handed (Comparison)	
Cases	75	15	90 Total Cases
Non-Cases	2,500	650	3,150 Total Non-Cases
	2,575	665	3,240 Grand Total

The columns are labeled with an index group and a comparison group. Notice that we have added margins to the table which are the sums of the rows and columns. You calculate the rate of substantiated cases among the right-handed persons by dividing 75 by 2,575; that of the left handed persons by dividing 15 by 665. You will see that there is a difference. It might look small, but looks can be deceiving. The rate among right-handed people is 0.029 (or 2.9%) while the rate in the left-handed group is 0.023 (or 2.3%). In this case, you see that your hypothesis looks like it was correct, right-handed people have a higher rate of child abuse than left-handed people, but how much higher? If you divide the rate for right-handed people (0.029) by the rate for left-handed people (0.023), you will calculate the rate ratio, 1.29.

The rate ratio has other names such as relative risk and risk ratio, but we will stick with rate ratio here. You may say that the difference between these two rates does not amount to much, but actually the interpretation is that the right-handed people have about a 30% elevated risk of being child abusers than the left-handed people. This is obtained by taking the value of 1.29 and subtracting 1 from it. The magnitude of the rate ratio for the period of time in question is calculated as the rate ratio minus 1, i.e., 1.29 – 1.00 = 0.29 or 29%. The statistical significance (whether the resulting rate ratio is likely to be due to chance) of this difference in rates can also be calculated.

Another statistic which is commonly reported in the literature is the odds ratio which is calculated by multiplying the cells in the table that are on the diagonals. For example, if you multiply 75 times 650 and divide that quantity by 15 times 2,500, you will get 1.3, about the same number as you got when you calculated the rate ratio. The odds ratio is a reasonable estimate of the rate ratio when the incidence of the events in question is low (less than 20%) and the prevalence of exposure is steady during the exposure period (Greenlander & Thomas, 1982). Both of these are always considered as estimates of the "true" effect, which is rarely known.

The four-fold table is frequently used in epidemiological research. In research where the investigator is interested in the effect of some exposure on the population, the headings for the columns of the four-fold table would be labeled "Exposed" and "Non-Exposed." They can be called anything you want as long as you note which is the index group (the group of interest) and which is the comparison group. A more complete discussion of these topics can be found in Rothman (1986).

References

Greenlander S & Thomas DC.(1982). On the need for the rare disease assumption in case control studies. *American Journal of Epidemiology*, *116*:547–553.

Rothman, K.J. *Modern Epidemiology*. 1986. Boston: Little, Brown and Company.

An Example of Chi Square: Computations and Degrees of Freedom

As we have previously noted, chi-square is a statistical test designed to answer questions about research data that exist in the form of frequencies (counts of event) rather than measurements or scores along some scale. Examples of possible frequency categories in which you may count occurrences are: male or female, yes or no, abuser or non-abuser, single or married, agree or disagree. Chi-square is a measure of association, not causality. It is important to note that neither the chi-square test nor other tests of association tell whether one event causes another. Therefore you should never make a causal inference based on a chi-square test. What does association mean? Association simply means related. It does not mean the degree of relationship, just that there is a relationship. The question to be answered by chi-square is whether or not frequencies (counts of events) observed in your sample differ significantly from chance. This comparison is made by comparing the distribution of your data with a theoretical or expected population frequency, the chi-square distribution.

The 2x2 table is a way of visualizing input for the chi-square test. How does the chi-square test work? Remember that chi-square tests the difference between existing, or observed frequencies and expected frequencies that are based on chance. Chi-square can also be described as a "goodness of fit" test, illustrating how obtained information (in the form of frequencies) differs from chance.

How do you calculate chi-square? Let's say that on Ft. Swampy during the last year there were 130 single active duty fathers and 1,000 married active duty fathers. Of the total of 1,130 fathers, seven (7) single fathers were child abusers and 120 married fathers were child abusers. A 2x2 table using this information can be constructed. (Cells of the table are assigned letter values to facilitate calculation.)

	Single	**Married**	
Child Abuser	[A] 7	[B] 120	127 (A + B)
Non-Child Abuser	[C] 123	[D] 880	1,003 (C + D)
	130 (A + C)	1,000 (B + D)	1,130 Grand Total

The research question is: Is there a significant difference between the proportions of child abusers who are single compared to those who are married? In other words, is there a relationship or association between child abuse and being either a single or married father? We can use a chi square test to determine if the association between fatherhood and child abuse is statistically significant at the 0.05 level. Basically, we are testing a null hypothesis that single fathers and married fathers have similar rates of abuse, i.e., that the frequencies for both groups are not statistically different. Using the following formula and the numbers from the above table, you can calculate chi square.

$$\chi^2 = \frac{[(AD\text{-}BC)]^2\ N}{(A+B)\ (C+D)(A+C)(B+D)}$$

$$= \frac{[(7)(880) - (120)(123)^2\ 1{,}130}{(127)(1{,}003)(1{,}000)(130)}$$

$$= \frac{83{,}574{,}800{,}000}{16{,}559{,}530{,}000}$$

$$\chi^2 = 5.05$$

You have determined that the chi square statistic is 5.05. To complete the statistical process and answer your research question you must compare the chi-square statistic to the chi-square table in the back of any statistics book (or your computer will give you the p-value). Remember that we are checking for statistical significance at the 0.05 level and we have one degree of freedom. From the 2x2 table, the degree of freedom is determined by multiplying (number of row categories minus one) times the (number of column categories minus one). Since we have two rows and two columns, our degree of freedom is (2-1) times (2-1) = 1. Any statistics book will show a value of 3.84 for the 0.05 level of statistical significance with one degree of freedom for the chi-square distribution. Because 5.05 is greater than 3.84 we can reject our null hypothesis that the two groups of fathers are the same. For this set of data, (which is fictional and created solely for this exercise), we can conclude that the abuse rates for married fathers is not only higher, but that there is a statistically significant difference in abuse rates between the single fathers and married fathers. Remember, however, that this in no way implies that being a married father causes one to be a child abuser.

More Information from the 2x2 Table

We continue to illustrate the types of information displayed in a 2x2 table. Specifically, we will show how to interpret the frequencies and percentages in each cell and in the margins of the table. To illustrate this, let's use a sample of 1,000 spouse abusers. The research question is: Is there is a relationship between the gender of offenders and incidents involving substance abuse? Below is the 2x2 table for this sample.

	Offenders			
	Male	Female		
Incidents Involving Substance Abuse 31%12.5%	[A] n=210 21% 84%	[B] n=40 4% 16%	Total = 250 25%	N Percent (Total) Row Percent Column Percent
Incidents Not Involving Substance Abuse	[C] n=470 47% 63% 64%	[D] n=280 28% 37% 87.5%	750 75%	
Total	680 68%	320 32%	1,000	

To answer this question, we can calculate the chi-square.

$$\chi^2 = \frac{[(AD-BC)]^2 N}{(A+B)(C+D)(A+C)(B+D)}$$

Here, the chi-square of 39.22 is statistically significant ($p<0.001$). Now we know there is an association between the gender of offenders and incidents involving substance abuse.

What other information can we gather from the table? In addition to the basic cell counts, or frequencies (N), each cell can provide three more categories of information. These are the overall percentage (the percentage of the total sample in that cell), the row percentage (the percentage of that cell's row total), and the column percentage (the percentage of that cell's column total). These percentages have been calculated for each cell. For example, the 210 males with incidents involving substance abuse constitute 21% of the total (1,000), 84% of the row total (250), and 31% of the column total (680).

Each of the percentages yields different information. Remember we are examining the association between the gender of offenders and incidents involving substance abuse. To illustrate this, we could report two different sets of percentages. Using the column percentages, (the percent of offenders with incidents involving substance abuse out of the total number of offenders for that column) we see that of the total offenders, regardless of gender, 25% (250) had incidents involving substance abuse. Distributed by gender, we see that of male offenders, 31% (210) had incidents involving substance abuse, and of female offenders, 12.5% (40) had incidents involving substance abuse. Using both frequencies and percentages, there are more male than female offenders with incidents involving substance abuse.

Remember, however, the differences between a frequency and a rate or percentage. The frequency is a count, or number of offenders, and a rate or percentage takes into account the size of the population. Had there been 900 males and 100 females in the sample, for example, we would get a different result: 23.3% (210) of the male offenders would have incidents involving substance abuse compared to 40% (40) of the female offenders. Using this example, females have a larger percentage of offenders with incidents involving substance abuse, even though there were more males with incidents involving substance abuse.

Using the row percentages to address the association between the gender of offenders and incidents involving substance abuse, we see that of all offenders regardless of substance involvement, 68% are males, and 32% are females. If we look at those incidents specifically involving substance abuse, we see that 210 (84%) are males compared to 40 (16%) who are females.

If there had not been an association between gender and incidents involving substance abuse, the gender distribution for incidents involving substance abuse would have equaled the gender distribution of the total sample. However, the distributions are not equal. There was a higher percentage of males with incidents involving substance abuse (84%) compared to the percentage of males in the total sample (68%). The females had a lower percentage of offenders with incidents involving substance abuse (16%) compared to the percentage of females in the total sample (32%).

Section 3

Measures of Association

In this article we present basic information about simple correlation, a measure of the degree of relationship between two variables, and a more complex type of correlation called cluster analysis. Correlation is a statistical measure of association. The simplest type of correlation involves two variables like height and weight. Cluster analysis is a more complex measure of association. It is a statistical technique that is used to organize (group) large amounts of data into meaningful structures. The example of cluster analysis is taken from the work of Dr. Ernest Jouriles.

Correlation

A correlation describes the degree of relationship between two variables. We will focus here on the most basic type of correlation in which the two variables are linearly (straight line) related and have a continuous scale of measurement. The correlation coefficient (also called the degree of correlation) is a single number that describes how closely related are two variables of interest.

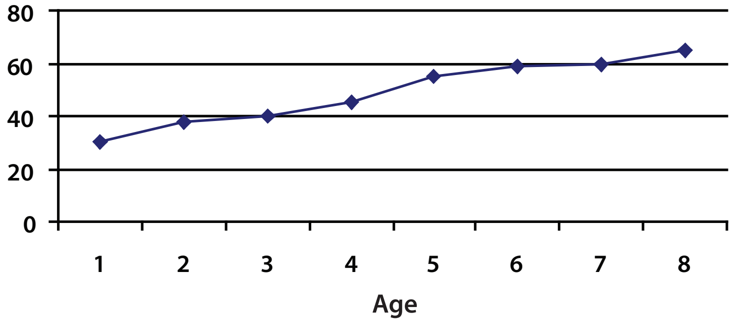
The correlation coefficient can vary from –1.00 to +1.00. If the value is greater than zero, it means that there is a positive correlation. That is, as one measure increases, the other increases. For example, as a child's age increases, the child's weight also increases. If the correlation coefficient is less than zero, it means that there is a negative correlation and that as one measure increases, the other decreases. For example, as the temperature increases, the number of inches of snow on the ground decreases. A perfect correlation (+1 or –1) indicates an exact correspondence between two measures. In behavioral science research, a correlation of 0.4 is considered a moderately strong correlation. A correlation coefficient of zero means there is no *linear* relationship between the two variables. However, there may be other relationships between the two measures. The type of statistic we are describing here does not address non-linear relationships.

As an example of a high positive correlation, let's examine the relationship between children's age and their weight.

Child	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Age	1	2	3	4	5	6	7	8
Weight in pounds	30	35	40	45	50	55	60	65

If you plot children's ages on the horizontal axis and their weight on the vertical axis, a fairly linear relationship is found between the two variables. We see that as one variable (age) increases, there is an increase in the other variable (weight).

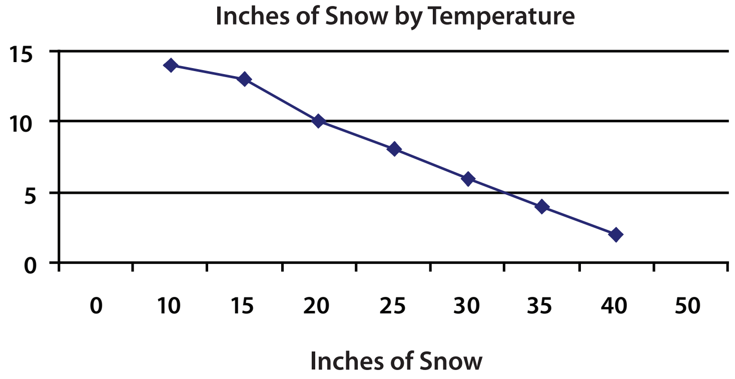
Weight of Child by Age



Let's examine the degree of relationship between the variables of temperature and inches of snow on the ground after a storm.

Snow Storm	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Temperature	10	15	20	25	30	35	40
Inches of snow	14	12	10	8	6	4	2

We can plot the temperature on the horizontal axis and the inches of snow on the vertical axis. You can see that as the temperature increases, the snow on the ground decreases. This is an example of a negative correlation that is also linear.



Correlation does not prove causation. Just because two variables change in a similar fashion (whether positively or negatively) does not necessarily mean that one causes the other.

Reference

Kranzler, G. & Moursund, J.(1995). *Statistics for the Terrified*. Upper Saddler River, NJ: Prentice Hall.

Cluster Analysis

Grouping ideas together is a basic cognitive process in which we all engage. Without some organizing strategy life would consist of an endless series of unique events leaving us without a mechanism to understand our world. Grouping can also lead to prediction, which may be helpful or unhelpful. An example of a helpful process is one in which we can use our cognitive organizing structures to better understand someone; an unhelpful one is stereotyping in which we perceive a person or an event in a rigid and inaccurate manner.

Cluster analysis is a type of statistical technique that is used to organize (group) large amounts of data into meaningful structures. It is an exploratory technique that can give numerical results, but cannot provide interpretation of those results. In other words, it provides a numerical (statistical) structure, but the investigator has to figure out what that structure means by observing which variables are grouped together. There are many different kinds of cluster analysis. The one selected depends of the type of analysis desired. For example, in exploratory research, the investigator can let a computer program determine the clusters. If the investigator is performing theoretical research and wants to see if the data conform to that theory, the number of clusters can be specified in advance.

We provide an example of cluster analysis in the work of Jouriles and colleagues (Grych, Jouriles, McDonald, Swank, McDonald, & Norwood, 2000). They used cluster analysis in their research as a technique to examine whether children of battered mothers differed on demographic variables, reports by children and mothers of interparental violence, children's reports of parent-child aggression, and children's perceptions and appraisals of interparental conflict. There are two main reasons why they used cluster analysis. First, it was exploratory research and, second, there was a large amount of data to be analyzed. There were 228 children between the ages of 8–14 in the study. Three measures were used for the cluster analysis: children's ratings of their internalizing behavior (anxiety and depression), children's ratings of their self-esteem, and mother's ratings of the children's externalizing behavior problems.

How did they use cluster analysis? They divided their sample into two groups and conducted separate cluster analyses on each in order to cross-validate their findings. In other words, they ran the cluster analysis on the first group and then ran the same analysis on the second group to see if they got the same results. This is a type of validation procedure that can be used when the sample is large enough to divide. Having these two separate groups saves the investigator from having to collect the same type of data twice. If the two samples yield similar results on the cluster analysis, they can be combined. If not, then the investigator must determine what was different about the two groups and conduct separate analyses.

The clustering procedure was based on children's reports of internalizing problems and self-esteem and mother's reports of externalizing behavior. A five-cluster solution provided the best description of the data based on the amount of variability accounted for (see Table). The more variability accounted for, the better the data fit the model. In other words, the procedure that accounts for the most variance leaves less information unclassified. After analyzing and comparing their two sub-samples of data, the results so were similar that they could combine the two and the final analysis was based on the total sample. They found the following clusters (represented on the left of the table) based on the pattern of results found by the procedure (represented on the right column of the table). The investigators named each pattern based on their examination of the results of the cluster analysis and one could read the table from right to left.

Table: Five Clusters of Adjustment of Children of Battered Women

Cluster	Percent of Variance	Pattern of Results
I. No significant maladjustment	31%	Low internalizing problems Low externalizing problems High self-esteem
II. Multi-problem externalizing	19%	Mainly externalizing problems High internalizing problems
III. Externalizing	21%	High externalizing problems High self-esteem
IV. Mild distress	18%	Slightly elevated internalizing problems
V. Multi-problem internalizing	11%	Elevated externalizing problems High levels of depression

The first and largest cluster was made up of children who were not exhibiting any signs of serious maladjustment (31%). Their scores were in the normal range of adjustment and none of the children or their mothers reported clinically significant problems. They also had the highest means on the self-esteem measure. The second group (19%) was labeled multi-problem externalizing. These children had elevated levels of both externalizing and internalizing problems, but more externalizing than internalizing scores. Only 9% of these children had mean internalizing scores above the clinical cutoff score. Thus, externalizing problems were predominant in this group. The third group (21%) had high externalizing scores, but none had high internalizing scores and their self-esteem was relatively high. The fourth group (18%) was labeled mild distress due to slightly elevated means on the internalizing scale and very low levels of externalizing problems. The fifth and smallest group (11%) was labeled multi-problem internalizing. They were distinguished by high levels of depression and somewhat elevated externalizing problems.

In conclusion, in this study cluster analysis demonstrated a reasonable method of organizing the varied patterns of adjustment of children exposed to and responding to interparental violence. The five patterns that emerged provided new information about such children's adjustment.

Reference

Grych JH, Jouriles EN, McDonald R, Swank PR, McDonald R, & Norwood WD. (2000). Patterns of adjustment among children of battered women. Journal of Clinical and Consulting Psychology, 68:84-94.

Section 4

Research Design

In this section, we begin with a theoretical article on how to consider the importance of the evidence presented in research. Six levels of evidence allow the researcher and, perhaps more importantly the reader, to consider the strength of the evidence based on the characteristics of the study. These six levels apply to results from clinical studies and suggest how much faith the reader can put in the conclusions. The lowest of these six levels is the untested treatment while the highest is that of randomized clinical trials. (The importance of evidence from randomized clinical trials is discussed in another article.) Next, we give an explanation of how to consider the meaning of statistical significance in hypothesis testing. We briefly discuss the use of one- and two-tailed tests, how to construct a hypothesis that can be tested statistically, levels of significance, number of subjects needed to conduct a study, and errors that can occur (Type I and Type II). Two related issues are confounding and bias. Both can occur in research design and lead to erroneous conclusions. Confounding occurs when the outcome you are studying is affected by a variable other than the one in which you are primarily interested and of whose existence you may be completely unaware. The confounding variable may mask or otherwise obscure the effect of the variable that you are attempting to study. Bias, on the other hand, occurs when there is a systematic problem in your study which will lead to an error in your conclusions. We discuss four different types of bias: selection, observational (or informational), recall, interviewer, and misclassification. The effects of bias are difficult to evaluate and often impossible to correct after a study is over. For this reason, it is very important to think carefully about all types of bias before you conduct your study and then to take steps to minimize its occurrence. The final article in this section discusses the psychometric properties of reliability, validity, and internal consistency in measuring child neglect.

Levels of Evidence

In statistics, one thinks about probability — the likelihood that a finding does not occur by chance. Conceptually, one thinks about research design issues such as what you measure, who your subjects are, what kinds of bias are likely to be encountered, and how you interpret your results. In terms of planning how to set up research, a brief discussion of how one conceptualizes the strength of evidence in research is presented.

It is important to be certain about what one says to the world in terms of published material, advice given to practitioners and clients, and information provided to the Army leadership. Often the results of scientific studies that you read about in the paper are presented as if they were true beyond question. Considering the strength of evidence is one way to increase your own knowledge about planning your own research and reading the results of the research of others.

Many FAP-related studies are clinical. That is, participants will be clients (or patients) of alleged or substantiated maltreatment in whom some effect is studied and, ideally, compared to non-clients. Examples are a prevention program targeting new parents to see if they abuse their children the future and studies of the effectiveness of some kind of intervention with either offenders or victims. When such studies are conducted, the investigator will want to know whether hypotheses are true or false and how sure one can be about the findings. In a book about the effectiveness of treatment for post-traumatic stress disorder, Foa, Keane, and Friedman (2000) listed six guidelines for evaluating treatment approaches used by clinicians. These guidelines are important in determining the strength of the evidence for the use of specific treatment procedures. They are general and can be applied to most clinical approaches in the mental health/social science field. We suggest that individuals planning or studying research consider how interventions or specific recommendations would be categorized in terms of the six guidelines.

Level l: The lowest of clinical evidence is a recently developed treatment that has not been clinically or experimentally tested.

Level 2: The treatment is based on long-standing practice by a small group of clinicians, but has not been experimentally evaluated.

Level 3: The treatment is based on long-standing and widespread clinical practice, but has not been subjected to experimental treatment.

Level 4: Evidence for this level is based on service and naturalistic and clinical observations that are compelling.

Level 5: Evidence is based on clinical research, but without randomized assignment to treatment groups or comparison groups.

Level 6: This level of evidence is based on well-controlled randomized clinical trials.

There are more considerations to be addressed in planning an evaluation study. The interpretation of the results of studies that fit into any of these levels of evidence is not independent of other considerations. Examples are the setting in which the research was performed, how the participants were selected or if they volunteered, the type of measurement used, statistical procedures, the size of the treatment effect observed, and whether the conclusions were based on the data gathered or were based on a generalization from the results.

Reference

Foa EB, Keane TM, & Friedman MJ. (2000). Guidelines for the treatment of PTSD. *Journal of Traumatic Stress, 13*:539–588.

Hypothesis Testing and Statistical Significance

Statistical significance is based on sample size and the variability of the data. Suppose you have two samples. The larger the sample sizes and the less variation in the data, the more likely it is that the two samples will be found significantly different by a test of statistical significance, say a t-test or a chi-square test. However, a finding might be statistically significant, but not very meaningful if there are large samples and very little variation.

While statistical significance may not be the whole story about the results of an experiment, it is still important. When statistical significance is lacking, one can say that there is a good chance that any differences observed are solely due to chance. In this article, we explore issues of statistical significance and discuss the types of error that can be made in hypothesis testing. Suppose you want to examine rates of postpartum depression in two groups of mothers, one group with no previous children and one with previous children. Each mother is given a scale to measure depression, and you find that one group has a higher average score than the other. You then perform a statistical test to see if the difference is statistically significant. However, before you do this, you have to decide on the hypothesis you are testing, the level of statistical significance that is acceptable, and the number of subjects necessary to perform a valid statistical test.

First, you state your hypothesis. In statistics, you try to reject the null hypothesis, usually written H_O. The null hypothesis is always that there is no difference between the means of your two groups. Your alternative hypothesis is that there is a difference between the means of the groups and is usually written as H_1. Why is it done this way? Because the probability associated with the test statistic (say a chi-square or a t-test) tells you the chances that you are wrong in rejecting the null hypothesis when it is true (e.g. finding a difference when there is none).

The level of significance depends on the amount of risk you are willing to take. There are two types of errors you can make. The first is called a Type I error and this is the error noted above, that you will reject the null hypothesis when it is true. This amount of error is stated as the significance level you are willing to accept. The level of 5% ($p<.05$) is the usual standard for statistical significance. This level of 5% means that you are willing to accept the risk that you are wrong in rejecting the null hypothesis 5% of the time, or 5 chances out of 100. The second kind of error is called the Type II error and this means failing to find a difference when one is actually there. Stated more formally, a Type II error means that you accept the null hypothesis (no difference between the groups) when it is false (there actually is a difference, but you do not detect the difference due to chance).

You have to decide if you are going to perform what is called a one-tailed test or a two-tailed test. If you predict the direction of the differences between your two groups, you will perform a one-tailed test. For example, if you think new mothers who already have children will have higher depression scores than new mothers with no previous children, you will have a one-tailed test. If you are willing to accept either alternative, i.e., that you are not hypothesizing which group will have higher depression score, then you will perform a two-tailed test. Most commonly, investigators perform two-tailed tests.

The final step is that of deciding how many subjects you will need to perform an adequate statistical test given the level of significance chosen, the variability of the data, and the size of the groups being compared. This is called calculating the power of a test. The power is actually defined as the probability of detecting a difference between the groups being compared when a difference really does exist. Numerically, it is 1 minus the Type II error. This step can get very complicated because a function can be plotted for power versus sample size for every hypothesized level of effect. This means that if you have some estimate of the variability of your data (or the size of the effect you are measuring) you will have a different power for your statistical test. Thus, the power of the test increases as the sample size increases. It should be noted that these measures may have a false rigor to them in that the selection of a level of significance and a sample size are somewhat arbitrary. All statistics are models of the effect you are investigating and such models will always have error. Controlling the amount of error is the bottom line in performing a good study.

Reference

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Sampling in Research Design

In previous articles we have discussed two issues that can affect the interpretation of research results: confounding and bias. In this article, we introduce the topic of sampling. Sampling is one of the most important concepts to consider when it is not feasible or realistic to measure an entire population. For example, in studying deployment, you cannot contact all deployed and non-deployed soldiers. So, you derive a sample from the population you want to study. The following are some possible questions you might ask in such a study:

(l) Are soldiers returning from deployment more prone to family violence than those who are not deployed?

(2) Is the length of deployment related to increased family violence?

(3) Is one category of family violence affected more than another category (spouse or child abuse; minor injury vs. major injury)?

(4) When are the effects of deployment most likely to be seen (before, during, or after deployment)?

Try to sample a population in such a way that you provide unbiased estimates of the effects you are measuring and sample in the most efficient manner in terms of time and money. A random sample, one in which every person in the population has an equal chance of being selected, is almost always the most desirable goal.

Construct your sample in such a way as to reduce error to the minimum. One way of reducing error is to stratify the population into groups of interest. For example, if you know your population contains more women than men, break up your population into women and men and take a sample from both groups. Stratification would ensure that your sample has enough male and female subjects. Stratification has been a problem in previous military research. Because of the small proportion of women relative to men in the active duty force, a random sample of the Army, in general, may not have enough female participants. Another common method of stratification is by age group.

What should you consider when reading a publication that uses a sampling methodology in its research design? Was the sample drawn randomly? Was the sample drawn in such a way to eliminate obvious biases? Is the sampling procedure as efficient as possible? Is the sample large enough to show the effect you are investigating?

The Statistical Concept of Confounding

In previous articles, we focused on ways to interpret data from studies in which there is only one variable. Usually, studies of social conditions and health involve the possible effects of many variables. For example, we often discuss several risk factors for abuse. The concept of confounding that is central to understanding the effects of several variables of an outcome.

Confounding occurs when the outcome you are studying is affected by a variable other than the one in which you are primarily interested. In other words, you believe that a particular variable (a possible cause) is responsible for the outcome you are studying, but another variable that you had not previously considered (or may not be able to do anything about) is affecting your outcome. This second variable, the confounder, may mask or otherwise obscure the effect of the variable of interest. A confounder is basically defined by two criteria: first, it is associated with the variable you believe is causing the effect, and, second, it is a possible independent cause (risk factor) of the outcome.

Since confounding is a difficult concept to understand, we present a simple illustration of a possible FAP research problem. Suppose you are studying the effect of a program to prevent child maltreatment by first time mothers. You first have to decide upon the risk factor that you would like to study. Suppose you believe that it is youthful age. But, there are additional risk factors possibly associated with young motherhood: lower income, a less mature marriage, separation from the family of origin, less adequate housing, or others that may be peculiar to your installation. Also suppose that you have a wide range of ages of new mothers available for your study. You decide to attempt to determine whether the younger new mothers are at greater risk for maltreating their children than the older new mothers. Your study consists of measuring the ages of all these mothers and then determining if age is related to the number of cases of child maltreatment. You find that younger first time mothers do have more child maltreatment incidents. You conclude that there is an association between mothers' age and child abuse. Your colleague, however, says, "Wait a minute. Some of these older women are the wives of senior NCOs and officers. They have enough money to hire extra help for a few weeks and they could buy more things for their children and probably did not have to worry about paying their bills. How do you know if the important factor in the number of child abuse incidents was not income? Maybe you should study the effect of family income on child abuse and not age."

You go back and look at your data and discover that the older women did have higher incomes. Therefore, it now appears that your study of the effect of mothers' age on child abuse may be confounded by family income. The confounder in your study was another possible risk factor (income) that is associated with age and is an independent predictor (negative) of child abuse. This satisfies the two criteria for a confounder as noted above. Now that you suspect that there is confounding, what do you do about it?

Confounding has to exist in the data you are studying. Just because the concept theoretically exists does not mean that it exists in your study. In order to determine if a confounder actually exists in your study, you have to statistically test for it.

You can control for confounding either by the design of your study or during the analysis. An example of controlling for confounding during the analysis is by stratification of the independent variable. You would analyze the low income women and high income women separately as if they were in different studies. If you found no statistical difference in their income, you would conclude that confounding was not present. If there were more child abuse incidents by low income women, you would then report that the relationship between age and child abuse depends on the effect of (the confounder) income.

For more information on confounding, see Rothman KL & Greenlander S. (1998). *Modem Epidemiology*. (2nd ed.). Lippincott-Raven.

The Statistical Concept of Bias

Confounding is part of a larger statistical concept called bias. Bias occurs when there is a systematic problem in your study which will lead to an error in your conclusions. There are many different types of bias. We will outline some of the more common ones.

Selection bias occurs when the subjects *selected* for the study do not represent the population you want to study. An example of when selection bias occurs is when the subjects in a study are selected from different populations. For example, Pope and Hudson (1995) describe a hypothetical study of eating disorder patients who attended a clinic for treatment and a control group recruited from the community. Those who participated in the study did not come from the same population. The ones with eating disorders may have sought therapy more than the controls. It is also possible to select a population of controls that are "supernormal," that is, free from occupational or psychiatric impairments. The remedy is to make sure that the subjects are selected by identical recruitment methods from the same population.

Observation bias, also known as *information bias,* occurs when information is incorrectly reported or concluded from the study participants. There are many types of observation bias.

Recall bias is one type which occurs when participants remember and report their experiences incorrectly. For example, if you are doing a study on alcohol involvement and its effects on spouse abuse, the participants who drank a lot prior to an incident of abuse may not remember much, and may report a mild incident of abuse compared to those who did not drink and could clearly remember what took place. Recall bias is also a type i*nformation bias,* which occurs when the investigator obtains information from one group differently than the other. In this case, if the interviewer knows which group the patients are from, there may be a tendency to give subtle cues or to ask more questions of the treatment group than the control group. The remedy here is to use the same information gathering tools and to not know which group the interviewee represents. Another form of information bias occurs when the subjects provide additional information beyond that which has been requested. In other words, the person with the problem may have reflected on the origin of the problem, read more material, or had more treatment than the person without the problem. This type of recall bias is difficult to remedy. One approach suggested by Pope and Hudson is to use only severe cases so that recall bias is minimized.

Interviewer bias occurs when the person conducting a study differentially collects, records, or interprets information from the subjects. For example, if an interviewer felt that children of single parents were more likely to be abused, he or she may ask the children of single parents more questions about being abused. Or, maybe the interviewer feels that alcohol use is related to spouse abuse, and records those people who drink as abusers, regardless of whether the abuse actually occurred.

Misclassification. Some degree of misclassification is present in almost all studies. Misclassification occurs when information on study participants is incorrect. You may record that subjects are married when they are single, or you may classify them as being involved in spouse abuse when they were not. Misclassification can occur simply by checking the wrong box on a form.

The effects of bias are difficult to evaluate and often impossible to correct after a study is over. For this reason, it is very important to think carefully about all types of bias before you conduct your study and then to take steps to minimize its occurrence. We close with a reminder to be critical of what you read. Pay attention to how bias and other confounding variables may affect the outcomes of a study.

Reference:

Pope & Hudson (1995). Does childhood sexual abuse cause adult psychiatric disorders: Essentials of methodology. *Journal of Psychiatry and the Law*, *23*:363–381.

Section 5

Interpretation of Results

This section is illustrated by giving examples of statistical procedures from recent research and discussing the meaning and significance of the statistical material presented in the article. By so doing, we hope to use concepts already presented and show how they are used in practice.

Mediators and Moderators

In many research articles in behavioral science, one often reads that an outcome is mediated or moderated by a third variable. These important terms are sometimes misused or used interchangeably, but they are very different concepts. Each has a different meaning for the understanding of research procedures and results. This article explains the differences.

A mediator is a factor that explains *how* or why the relationship exists (Baron & Kenny, 1986). In order for a factor to be a mediator, it must lie on the pathway between the independent variable (the factor you are interested in studying) and the dependent variable (the outcome). In order to be a mediator, a variable must demonstrate a significant degree of relationship between the independent and the dependent variable. If no relationship exists, then the hypothesized mediator does not lie on the causal path and hence cannot be a mediator.

To illustrate, we use an example from Buckner, Bassuk, & Beardslee (2004) who examined the association between exposure to violence and mental health in poor children. They found that children exposed to violence experienced more mental health symptoms than those who had not been exposed (the direct relationship). To help explain this relationship, they investigated four factors as possible mediators of violence and mental health symptoms: perceptions of environmental danger, locus of control, self-esteem, and emotional regulation. The authors found that exposure to violence led to lower self-esteem and a higher perception of danger, both of which, in turn, led to internalizing symptoms and poor mental health. Therefore, self-esteem and perceptions of danger are mediators in the relationship between exposure to violence and mental health; they help explain why exposure to violence is related to poor mental health.

In contrast, moderators explain *what* or in what subgroups certain relationships exist. In other words, moderators help us understand if there are certain characteristics of people or environments that make the relationship between the independent variable and the outcome stronger or weaker. A moderator may affect the direction or the strength of the relationship of interest. A moderating variable should have little or no statistical relationship to either the independent or the dependent variable.

Gender is often a moderator. In the Buckner study (2004) mentioned above, there was a relationship between exposure to violence and mental health symptoms (internalizing symptoms such as anxiety, depression, and somatic complaints) and it was stronger for girls than for boys. Thus, gender was a moderating variable in the relationship between exposure to violence and mental health symptoms. It affected one group (girls), but not the other (boys).

The differences between mediators and moderators are more complex than this presentation. We have only highlighted the differences. The reader is referred to Baron and Kenny (1986) for detailed descriptions of these concepts and to Buckner, Bassuk, and Beardslee (2004) for more detail on their analyses of moderating and mediating variables in the association between children's exposure to violence and mental health symptoms. Overall, moderators and mediators help us understand relationships, and have important implications for the development of prevention and treatment interventions.

References

Buckner JC, Bassuk EL, & Beardslee WR. (2004.) Exposure to violence and low-income children's mental health: direct, moderated, and mediated relations. *American Journal of Orthopsychiatry, 74*:413–423.

Baron R & Kenny D. (1986.) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*:1173–1182

Prospective and Retrospective Approaches to Child Maltreatment Research

The terms prospective and retrospective are used to describe two types of research design. Understanding a study has been designed often requires a detailed examination of the research methods. Terms, such as longitudinal, case-control, cross-sectional, and cohort, are apt to be confusing in the context of a prospective or retrospective design. In this article, we will clarify the basic distinction between prospective and retrospective designs and show their relation to other terms. We will then present some views of two groups of researchers on the benefits and limitations of prospective and retrospective studies in child maltreatment research.

Let's consider how we think about the relation between cause and effect. One way is to attempt to relate an event, which is called an exposure (such as childhood maltreatment) to an outcome, such as an adult illness or a symptom. Research design requires that both the exposure and the outcome be measured, that their temporal sequence is reasonable (e.g., the outcome cannot occur before the exposure), that it is possible to analyze the relationship between the exposure and the outcome, and that the results are plausible (e.g., conform to a theory or fit in with previous findings). The most important distinction between prospective and retrospective studies is that in a prospective study measures of exposure are taken before the outcome has occurred while in a retrospective study the measure of exposure is taken after the exposure; that is, retrospectively (e.g., looking backwards). In a prospective study, a group of children who have not been exposed to maltreatment are identified and followed over time. In a retrospective study a group of children, some of whom have already been exposed, are identified and measures of exposure are taken after the outcome has occurred. For example, children who have been maltreated are assessed for their history to investigate variables that were associated with maltreatment such as low birth weight. Another way of stating this distinction is that the two methods differ in the timing of subject (case) identification. Prospective studies identify individuals or study groups from a population that will be followed for a period of time to determine the outcome. Retrospective studies take the outcome and then, looking back, determine what significant events occurred prior to the outcome (Greenlander & Rothman, 1998).

Other distinctions add complexity to the descriptions of both prospective and retrospective study methods. Two additional terms, cohort and case-control, are important. Frequently these are misidentified as describing prospective and retrospective designs, respectively. In cohort studies, participants are selected according to their exposure status (e.g., soldiers who have not yet deployed); in case-control studies, participants are selected based on their outcome status (e.g., all soldiers with posttraumatic stress symptoms after return from deployment) (Greenlander & Rothman, 1998). However, both cohort and case-control studies can be prospective or retrospective.In longitudinal studies, repeated measures are taken on the same persons and they are identified so they can be re-tested. A cross-sectional study is like a snap-shot in that measurement of exposure and outcome occurs only once and at the same time. A series of cross-sectional studies can be performed on a population to describe changes in the population over time, but usually the subject cannot be identified and linked to other information. An experiment is always a prospective cohort study because subjects are selected and assigned to groups and the investigator then waits for the outcome to occur (Greenlander & Rothman, 1998).

In a recent issue of *Child Abuse & Neglect* two groups of researchers offered comments on some advantages and disadvantages of both prospective and retrospective studies in maltreatment research (Widom, Raphael, & DuMont, 2004; Kendall-Tackett & Becker-Blease, 2004). Each type of study (prospective and retrospective) has its advantages and drawbacks. One should not assume that prospective is necessarily better than retrospective.

Problems of Retrospective Studies In Child Maltreatment Research

Accuracy of information. One of the problems of retrospective studies involving self-reports is whether the information provided is accurate. Why might such information be inaccurate? What a person remembers from childhood might be dependent on what the person has been told. There is a considerable body of maltreatment literature showing an unacceptable level of validity (accuracy) of self-reported (retrospective) childhood experience. Among the other reasons for such lack of validity are lack of rapport with the interviewer, a desire to protect parents or other persons, and a desire to forget or deny the past. Additionally, in retrospective reporting it is almost impossible to determine the extent of false positive responders, persons who say that an event happened when, in fact, it did not happen (Widom, Raphael, & DuMont, 2004).

Types and Sources of Bias

Recall bias can cause errors in retrospective reports. Recall bias occurs when persons report exposure information after learning that they have the outcome in question (Greenlander & Rothman, 1998). Other examples of why people may be more likely to report early experiences in a negative way (recall bias) are poor health, negative mood, and other factors in the current life of the individual such as depression, substance abuse, and life satisfaction (Widom, Raphael, & DuMont, 2004).

Sampling bias can occur in retrospective studies. It may be difficult to obtain a sample of the most representative population for the problem one wishes to study. For example, different data are usually obtained from persons visiting a doctor than from those in a women's shelter or from college students. Each of these will be biased in the direction of the problems presented by the respondents in each of these situations and can be representative only of that population (Widom, Raphael, & DuMont, 2004).

Investigating causality versus risk. In retrospective reports there is little chance of examining causal relationships between exposure and outcome whereas this is more likely in prospective studies. Whether outcomes are directly or indirectly related to the exposure will be difficult to tease out, but prospective studies at least allow the investigator to learn the temporal sequence of events following the exposure and other adverse events. While retrospective studies may not allow one to draw conclusions about causality, they can suggest possible risk factors for the outcomes (Widom, Raphael, & DuMont, 2004).

Problems of Prospective Studies In Child Maltreatment Research

Identification of participants for the research. There are many problems in identifying groups of children to follow in prospective studies. In one type of prospective study, an investigator would follow a group of children and later identify those children who are maltreated and those who are not. However, it is hard to identify maltreated children. Prospective designs will probably miss many victims of childhood maltreatment whose maltreatment was never reported to authorities. When victims are identified, reporting to authorities is mandatory. Investigators cannot simply identify and follow them without taking into account the effect of their identification and intervention or non-intervention. Finally, persons who were identified as maltreated children are probably not representative of maltreatment survivors as a whole. Thus, prospective and retrospective studies are likely to identify separate subgroups for study (Kendall-Tackett & Becker-Blease, 2004).

Severity of abuse. Unreported abuse may be more severe. Abuse may be more severe when unreported due to the belief that when abuse is identified it is more likely to stop; when it goes unreported it can continue and even escalate becoming more frequent and more severe. When maltreatment goes unreported, there can be other associated outcomes such as shame and isolation that can result in different outcomes such as more symptoms (Kendall-Tackett & Becker-Blease, 2004).

Costs. Prospective studies are very expensive when the investigator seeks to study low frequency events. For example, one must follow 100 subjects to find one case if the rate of occurrence is 1%.

The statistical issues involved in the distinctions presented here are more complex than our presentation here. However, our purpose is to present the broad outlines of prospective and retrospective research designs and to apply them to child maltreatment research and practice.

References

Greenlander S. & Rothman K. (1998). *Modern Epidemiology*. Philadelphia: Lippincott-Raven, Inc (p. 74–75, 114).

Widom CS, Raphael KG, & DuMont KA. (2004). The case for prospective longitudinal studies in child maltreatment reseach: commentary on Dube, Williamson, Thompson, Felitti, & Anda. *Child Abuse & Neglect*, *28*:715–722.

Kendall-Tackett K & Becker-Blease K.(2004). The importance of retrospective findings in child maltreatment research. *Child Abuse & Neglect*, *28*:723–727.

Defining and Measuring Child Neglect

Defining and measuring child neglect is challenging. Increased knowledge of measures of neglect can aid the Army Family Advocacy Program in its prevention and treatment missions. Sound, empirically based assessments are needed. In this section, we give some examples of neglect definitions and measures that have been developed and used in research and clinically.

Straus and Kantor (2005) suggest a definition of neglect, provide a conceptual analysis of that definition, and identify principles, criteria, and problems in creating measures of neglect. Their definition highlights the *neglectful behaviors of a caregiver,* failures to meet the developmental needs of a child. Similarly, Dubowitz (2005) distinguishes between failures to meet needs in contrast to inflicting harm. Its causes and motives are different.

Kantor and Straus, at the Family Research Laboratory at the University of New Hampshire, have developed a number of measures of child neglect. One of these is the Multidimensional Neglectful Behavior Scale-Child Report (MNBS-CR) (Kantor et al., 2004). It measures four primary domains of neglectful behavior: emotional, cognitive, supervision and physical neglect. Good psychometric properties were demonstrated in their validation samples.

The term psychometric properties refers to measures that have been obtained in the development of an instrument. Generally, at a minimum, these include measures of reliability and validity. Reliability and validity are two basic concepts in test development and measurement. There are many types of both reliability and validity. In general, reliability refers to the consistency of a measurement. Test-retest reliability is the degree of agreement achieved when a measure is given on two different occasions. For example, if intelligence is measured on two separate occasions under the same circumstances, the results should be very similar. Inter-rater reliability is a measure of the degree of agreement of two or more persons rating the same event. An example of inter-rater reliability is the degree of agreement between persons judging candidates for a job. Validity, on the other hand, is a different concept. In general, validity indicates the degree to which you are measuring the concept that you are attempting to measure. There are several types of validity. Concurrent validity is the degree of agreement of a new measure, say a test, with one that has already been validated. Predictive validity indicates how well a test predicts some criterion. For example, in maltreatment research, we would like to have a measure with high predictive validity for recidivism. That is, it would predict recidivism risk with high probability. High predictive validity is generally required for clinical use.

Internal consistency reliability is another psychometric property. Straus and Kantor's measurement development procedures included a clinical sample of 144 children, ages 6–15 and a comparison sample of 87 children. The full version of the MNBS-CR had high internal consistency reliability among both the younger (alpha=.66) and older children (alpha=.94) with neglect concerns. Alpha is a measure of the internal consistency of a scale. Internal consistency measures the extent to which scale items correlate with each other. The higher the value of alpha, the more the items measure the same idea and the higher is the internal consistency. Scores above 0.60 indicate reasonable internal reliability. For more information on scales and measures, see http://www2.chass.ncsu.edu/garson/PA765/standard.htm].

Correlations among the MNBS-CR subscales ranged from moderate to high indicating overlap between the subscales. Correlational analyses between the total neglect scores and child outcomes provide some support for construct validity of the MNBS-CR. In addition, analyses of the relationship between MNBS-CR reports and caretaker reports were conducted to address construct and predictive validity. (The MNBS-CR scales should be used only with permission of Straus and Kantor. Contact the authors at: http://www.unh.edu/frl/unpubpap.htm)

Another instrument that focuses upon the measurement of child neglect is the Child Neglect Index (CNI) (Trocme, 1996). It was designed as a substantiation tool for child welfare practitioners and researchers to easily measure the type and severity of child neglect. Trocme's definition of neglect is based on criteria used by child welfare workers and, accordingly, reflects a more legal than clinical approach. In contrast to Straus and Kantor's conceptualization of neglect, the CNI defines neglect in terms of the different forms of physical or emotional harm that is seen in neglected children.

The CNI is a single page instrument including the following six scales: supervision, nutrition, clothing and hygiene, physical health care, mental health care, and developmental/educational care. For all scales an inadequate or neglect rating requires evidence of impairment or harm or exposure to situations that could cause harm.

The CNI was field tested in a large welfare agency on 127 consecutive intake investigations. Two scales, psychological care and developmental care, were correlated above .50. Test-retest reliability was assessed by the completion of the CNI by the intake workers twice within two weeks. Test-retest reliability scores for each scale were acceptable, with a range from .83 (developmental/educational care) to .91 (supervision). Interrater reliability scores on individual scales ranged from .69 to .95 with a mean of .79. Validity and reliability of the CNI compare favorably to longer and more detailed measures of child neglect.

Straus and Kantor (2005) give additional helpful information on the conceptualization and measurement of neglect in a recent article. They discuss neglect definitions, principles and criteria for the measurement of neglect separately from harm, various measures of neglectful behavior and their psychometric properties, and implications for research and practice. Although there are differences in the measurement of child neglect, there continues to be an ongoing need for interventions and supportive services for neglected children and neglectful parents. Enhancing our knowledge and understanding of measurement characteristics associated with the study of child neglect should significantly contribute to this important task.

References

Kantor GK, Holt MK, Mebert CJ, Straus MA, Drach KM, Ricci LR, MacAllum CA, & Brown W. (2004). Development and preliminary psychometric properties of the Multidimensioonal Neglectful Behavior Scale—Child Report. *Child Maltreatment, 9*:409–428.

Straus MA & Kantor GK. (2005). Definition and measurement of neglectful behavior: some principles and guidelines. *Child Abuse & Neglect, 29*:19–29.

Trocme N. (1996). Development and preliminary evaluation of the Ontario Child Neglect Index. *Child Maltreatment, 1*:145–155.

Effect Size Measures

We have previously discussed the issue of statistical significance and its interpretation. These discussions focused on two major points. First, that statistical significance is based on two factors in the data: sample size (the larger the sample the greater the likelihood of significance) and the variability (the less the variation the more the likelihood of significance). Second, that a finding may be statistically significant, but not meaningful, particularly in clinical work where it is important to be able to have confidence that your work will make a difference in a person's life, not just in a statistical table. How does one tell if some procedure, say some form of therapy, is worthwhile (a non-statistical term)? We would like to re-visit this topic of meaningfulness with an explanation of more terms that one may hear in presentations at meetings or see in the scientific literature, particularly in conjunction with evaluations of the effectiveness of therapies.

This discussion is presented in service of the Army FAP's attempt to develop outcome measures. One reason for developing outcome measures is to find the most effective treatment for victims of family violence and for those who offend. The last few years have seen great progress in psychotherapy research, although it has not been without controversy. The thrust of research in psychotherapy for the last few years has been toward identifying empirically supported therapies, finding which treatments have been shown via research to work with which patients? Three terms are frequently used in the evaluation of psychological treatment research literature (Chambless & Hollon, 1998): *efficacy* (does the treatment work in a controlled setting?), *effectiveness* (does the treatment work in actual clinical practice?), and *efficiency* (is the treatment cost-effective with regard to other interventions?).

One method of attempting to provide answers to each of the three questions above is through the use of a measure of effect size. Many professional journals have recently provided editorials about how to interpret effect size as well as other statistical topics such as tests of statistical significance, p-values, and confidence intervals. For example McClure (1999) recommended looking at effect sizes in addition to traditional tests of significance. Although effect size is not a new concept in statistics, it may be seen more often now due to the increased number of studies of psychological treatments.
In the most general terms, an *effect* is a difference in some phenomenon between two populations that differ on a characteristic, sometime called an exposure. For example, using standardized measures, you may find a difference in depression (the phenomenon of interest) in two groups of people, those with a history of abuse (the exposure) and those with no such history. This effect may be absolute (the difference between the means of the two groups on the standardized measure) or it may be relative (the mean of the exposed group divided by the mean of the unexposed group). So, one measure of an effect size is the magnitude of this difference. That is the most basic type of effect size. Another measure of effect size is that derived from a meta-analysis. A meta-analysis is a type of statistical procedure to obtain a statistic giving a measure of the results from a number of studies to determine if a phenomenon is present based on aggregated data. (Maxwell Smart might have said: "If you don't believe one study, would you believe five?") This type of measure is often presented at meetings when one therapy is compared with another or there are multiple comparisons on the same type of therapy.

Here is a general example. Investigators use the same measure of depression, but in different groups (samples) of people. The investigators might obtain the mean differences in mean depression scores and divide by the pooled standard deviation (square root of the variance) of the individual studies. (See Senra, 1995, for a formula for a pooled standard deviation.) The effect size is thus expressed in standard deviation units. An example of effect sizes from a multiple therapy groups is shown in a recent study comparing cognitive versus behavior therapy in the treatment of obsessive-compulsive disorder (McLean et al., 2001). Over separate groups, an effect size of 1.62 was found for exposure therapy and 0.98 for cognitive-behavior therapy. (These effect size statistics were calculated by subtracting the mean of the Yale-Brown Obsessive-Compulsive Scale for the wait-list control group from the mean of either of the treatment groups and dividing by the pooled standard deviation.) Cohen (1988) reviews the concept of effect size and notes that (1) it does not imply causality, and (2) it is a measure of the degree to which the phenomenon under study is present in the population that was investigated. In other words, no effect size, no effect. Cohen gives guidelines for effect sizes based on the d-statistic (difference between means divided by the standard deviation): small, 0.20; medium, 0.50; large 0.80. These guidelines are often repeated; however, (1) they are arbitrary, and (2) the clinical relevance of a treatment effect cannot be deduced from an effect size (Scholten, de Beurs, & Bouter, 1999). Greenlander (1998, p. 672) advises avoiding it because expressing effects in standard deviation units can yield spurious results in which identical results can be made to appear different or even reverse the order of the strength of results. He advocates expressing effects in a substantively meaningful unit that is uniform across studies, not in standard deviation units. The reason for this rather strong warning is because the variability across studies is almost never uniform and it is notoriously true in the behavioral sciences. Thus, standard deviation units are probably useless in most cases for comparing effect sizes.

Scholten, De Beurs and Bouter (1999) using fictitious data demonstrated how an attempt to transform effect sizes of decreases in blood pressure due to a new drug produced incorrect and confusing results. They noted that translation of effect sizes into clinically meaningful units is hazardous and that assessment of a treatment effect using only effect sizes is challenging.

So, how do you use the concept of effect size when analyzing the results of a study? First, it is difficult (unless you work consistently in this area of statistics) to make an intuitive interpretation of what an effect size means. We advise looking at the magnitude of changes in the basic measures, such as the differences between the means of the experimental and control (exposed versus unexposed) groups and at the magnitude of variability on the actual measures of interest. If these figures make sense, you probably can derive your own estimate of whether one is greater or of more importance to you than the other and an approximate effect size. Second, look at the way that effect size has been computed and make sure (1) it has been calculated correctly and (2) you understand what it means. As we have pointed out, there is more than one type of effect size, it may mean something different than you think, and it may not be calculated correctly. Third, make sure you distinguish effect sizes from other statistics such as p-values, odds ratios, and confidence limits. They do not all have the same interpretation.

References

Chambliss DL & Hollon SD. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*:7–18.

Cohen J. (1988). *Statistical power analysis in the behavioral sciences.* Hillsdale, NJ: Erlbaum. Greenlander S. (1998). Meta-analysis. In Rothman KJ & Greenlander S. *Modern Epidemiology* (Second Edition). Philadelphia, PA: Lippincott-Raven.

McClure PW. (1999). Determining the significance of significance: p-values, effect size, and clinical judgment. *Journal of Hand Therapy,* 41.

McLean PD, Whittal ML, Söchting I, Koch WJ, Paterson R, Thordson DS, Taylor S, & Anderson KW. (2001). Cognitive versus behavior therapy in the group treatment of obsessive-compulsive disorder. *Journal of Consulting and Clinical Psychology, 69*:205–214.

Scholten RJPM, de Beurs, & Bouter LM. (1999.) From effect size into number need to treat. *Lancet, 354*:598.

Senra (1995.) Measures of treatment outcome of depression: an effect size comparison. *Psychological Reports, 76*:187–192.

Rate Ratios and Confidence Intervals

These are two very common statistics used in reporting estimates of the size (or strength) of a finding [rate ratio] and the range of possible values of that estimate [confidence interval]. A rate ratio is the ratio of one rate to another and is reported as a single number whose value represents the strength of the finding. If there were no difference between two groups, the value of the rate ratio would be one. Rumm et al. (2000) reported a rate ratio of 2.0 for child abuse among families with an identified incident of child abuse, twice the value for families without an identified incident of spouse abuse.

The rate ratio is a point value. That is, it is a single data point. If one wants to know possible ranges for this value, a confidence interval is computed. The confidence interval represents the range of possible values that the rate ratio could take given a stated probability value, usually 95%. The stated probability value for the confidence interval is selected by the investigator to represent a reasonable limit that could incorporate possible values of the rate ratio. The width of the confidence interval depends on the variability of the data (the more variability in the data, the wider the interval) and the size of the interval selected.

Rate ratios and confidence intervals are ways of presenting information about the results of a study in addition to the value of statistical significance based on a test such as a t-test, a chi-square test, or others

Reference

Rumm PD, Cummings P, Krauss MR, Bell MA, & Rivara FP.(2000). *Child Abuse & Neglect, 11*:1375–1381.

Statistical Concepts in Risk Assessment

Risk assessment is one of the most important concepts in family maltreatment practice and research. Police and emergency workers conduct risk assessment in situations in which a judgment is made on the spot about a victim's safety. Risk assessment also is conducted statistically, in which predictions are made about safety, but using an instrument. The latter is the context for this article. The material presented here is applicable to most test instruments in which the user seeks a prediction that will be useful, such as the results of a screening test for disease. The major concepts that must be considered for instrument development and hence fall under this section of research design. These concepts are test sensitivity, specificity, positive predictive value, and negative predictive value. They apply to screening tests that predict whether a condition is likely to exist. Risk assessment research requires an understanding of the following terms:

- *Reliability*. A similar outcome is obtained if the measures are taken again under the same or similar circumstances.
- *Validity.* This term generally refers to the fact that the measure reflects the concept that is sought. Does a scale that purports to measure depression provide a measure of whether a person is depressed, or how depressed they are?
- *Sensitivity.* The ability of a test to identify if a person has the condition. Sensitivity is calculated by taking the number of true positives (people with the outcome) and dividing by the sum of true positives plus false negatives. (True positives are people with the outcome who are correctly identified by the test. False negatives are people who actually have the outcome, but are not detected by the test.)
- *Specificity.* The ability of a test to identify if a person does not have the outcome. It is calculated by taking the number of true negatives (people

without the outcome) and dividing by the sum of true negatives plus false positives. (True negatives are people who are correctly identified by the test as not having the outcome. False positives are people who do not actually have the outcome, but are seen by the test as positive for the condition).

- *Positive predictive value.* The likelihood that a person with a positive test has the outcome. It is the probability that someone with a positive test actually has the outcome. It is calculated by taking the number of true positives and dividing by the sum of true positives and false positives.
- *Negative predictive value.* The likelihood that a person with a negative test does not have the outcome. It is the probability that someone with a negative test does not have the outcome. It is calculated by taking the number of true negatives and dividing by the sum of true negatives and false negatives.

For a test result to be determined as positive or negative, there must be a set criterion point. Those above or below this criterion point are judged as positive or negative for the outcome of the test. If the criterion is set high, there will be more false negatives (people who have the outcome, but are not selected for the outcome). Alternatively, if the cut point (criterion) is set low, there will be more false positives (people who do not have the outcome, but are said by the test to be positive).

There are other statistical procedures that contribute to the determination of whether a test is useful or not such as the prevalence of the outcome in the population. (Is it a rare or a common condition?)

If one is predicting risk by perpetrators, what are the implications of having false positives and false negatives? False positives incorrectly identify people as likely to commit a violent event. False negatives fail to identify a person who is likely to commit a violent event. Sensitivity and specificity often (or usually) work in opposite directions. If sensitivity is high (people with the condition are identified), the specificity is usually somewhat lower (people who do not have the condition are not ruled out). It is important to have high sensitivity when you do not want to miss correctly predicting the outcome when it could be harmful or lethal. High sensitivity is especially hard to achieve in a population with a low prevalence of the outcome. This discussion urges the reader to be wary of claims for instruments and methods and insist on reviewing supporting data prior to use.

Screening for Dangerousness

Dr. Jacquelyn Campbell and other investigators in 1985 and 1988 (Campbell, et al., 2003) developed the Danger Assessment (DA) instrument to assist abused women in estimating their risk of homicide. (See http://www.dangerassessment.com.) A second purpose of the DA is to assist persons who work with domestic violence victims, such as police, advocates, and health care professionals, in measuring and warning women of their danger level. We make no evaluation of the DA, but use it as an example of some of the statistical properties of screening instruments for purposes of illustration.

The DA is conducted in two parts. First, the severity and frequency of assault is measured by presenting the woman with a calendar of the past year. She is asked to mark the approximate days when physically abusive incidents occurred and to rank the severity of the incident on a 1 to 5 scale where 1 is the least severe. The second part of the DA is a 20-item yes/no response format of risk factors associated with intimate partner homicide. Examples include "Has the physical violence increased in frequency over the past year?" and "Does he ever try to choke you?" The DA is scored by counting the "Yes" responses.

Among the statistics presented on the DA website are estimates of its reliability, validity, sensitivity, specificity, cutoff scores, and the receiver operating curve (ROC) analysis. Reliability statistics for the DA are provided for internal consistency (how well each item relates independently to the rest of the items on the scale) and test-retest (the correlation between two or more administrations of the same scale). Validity statistics are given for discriminant construct group validity (how well the instrument discriminates between groups) and convergent construct validity (how well the measures that should be related are related. Convergent validity means that different measures converge on the construct that you measure. Predictive validity is the ability of an instrument to predict what it is supposed to predict. The ROC is a graphical representation of test characteristics, such as the sensitivity and specificity, used in the evaluation of cutoff points for screening tests.

The effects of different cutoff scores on prediction using the DA are also presented. At this point, the statistical concepts become more difficult to understand. There are many ways to describe measures of how well a screening test actually works. A cutoff score on a screening test balances at least two essential concepts in prediction: sensitivity and specificity. In this example, sensitivity is the ability of a test to correctly identify the persons who are in danger. Specificity is the ability of the test to correctly identify persons who are not in danger. The investigator can set a cutoff score to select persons correctly screened (sensitivity) and eliminate those who should not be selected by the test (specificity). However, since tests are not perfect and do not represent reality, there will always be false positives and false negatives.

If sensitivity is high (i.e., you correctly identify the people in danger, then specificity is often low. If specificity is high (i.e., you correctly identify the people who are not in danger) then sensitivity is often low. For example, for a cutoff of 4 on the DA, about 80% of those in danger were correctly identified (sensitivity), but only about 40% of those who were not in danger were correctly identified. At a cutoff of 7, 58% of those who were in danger were correctly identified and 87% who were not in danger were correctly identified. It was noted that the sensitivity of 58% was worrisome because an additional 42% of the women in danger were not identified. The authors of the psychometric data page provide more information about an improved scoring system that involves a weighted score that correctly identifies 90.8% of the cases.

The validity of the DA has not been specifically established for any military population. As a result, some of the items that were particularly predictive in the civilian population may not have the same predictive power in the military (Campbell, Webster, Koziol-McLain, et al., 2003). For example, gun ownership was important in the 12-city study. However, in the military many soldiers collect guns and this item may not have the same predictive power as was found in a civilian population. Unemployment of the perpetrator was also predictive of homicide, but this item would not apply to an active duty military perpetrator population since all are employed.

Screening is a complex undertaking. Those persons contemplating using screening instruments should understand the concepts of screening as well as the implications of false positives and false negatives.

Reference

Campbell, J. C., Webster, D., Koziol-McLain, J., Block CR, Campbell, D., Curry, MA, Gary, F, Sachs, C. Sharps, PW, Wilt, S., Manganello, J., Xu, X. (2003). Risk factors for femicide in abusive relationships: Results from a multi-site case control study. *American Journal of Public Health*, 93, 1089–1097.

Gold Standard, Randomized Trials, Effectiveness, and Efficacy: What Do These Terms Mean?

What is a gold standard?

The term *gold standard*, in practice and research, denotes the highest possible level of value and is used for the purpose of comparison. Gold standard comes from the field of economics in which gold once represented (and sometimes still does) the monetary value of a country. In scientific research and practice, the gold standard is used to convey that which the researcher or practitioner holds up as the best means of measurement. While an autopsy might be considered the gold standard for findings related to pathology, an x-ray, MRI or CAT scan would be a radiologist's gold standard for diagnosis. In other words, one person's gold standard might not be another's! A gold standard is not infallible, just the best that is known.

What is a randomized trial?

A *randomized trial* (sometimes also called a randomized clinical trial or a randomized controlled trial) is used in research in which the investigator wishes to test the effect of an intervention (such as a new psychotherapy or new medication). The term randomized trial comes from the fact that participants (people or families or whatever unit you wish to study) have an equal chance of being assigned (i.e., random assignment) to different groups. Randomization (assignment to one of the groups to be tested) is a very important process and usually involves the use of a computer program, a random number table, or other mathematical procedure. The importance of randomization is to ensure that the two (or more) test groups are equivalent — having no systematic differences except for the intervention. The two or more groups are then used to compare different treatments, different amounts of some treatment, one treatment with another treatment, or with no treatment. All groups are given the same outcome measures to determine whether one treatment is better than another or is better than no treatment. While the perfect randomized trial may be difficult to achieve in practice, it is still generally the only accepted procedure that is recognized and approved by the FDA and other government agencies as demonstrating that a treatment works. It may not always be possible to perform a randomized trial for ethical or other reasons. For a definition of randomized clinical trials and other clinical terms see *http://www.cancer.gov/dictionary.*

What is the difference between an effectiveness study and an efficacy study?

A recently published study of the effects of home visiting (Duggan et al., 2004) is an effectiveness study. An *effectiveness* study is one in which the procedure (typically psychotherapy, but in this case home visiting) is tested *as it is actually performed in the field.* The efficacy study is conducted very differently. Efficacy studies are used to test if a specific procedure has any therapeutic value under ideal conditions. In an efficacy study, as many variables as possible are controlled. The experiment is done in a more rigorous manner with substantiated exclusion and inclusion criteria resulting in highly selected participants. In both types, therapeutic procedures are standardized and made explicit, usually by writing a treatment manual. (This is called manualization of the therapy.) The therapist then follows the treatment manual, which indicates what is to be done in each session and the number of treatment sessions. In an efficacy study, the fidelity of the therapist (how well he or she is following the procedure) is documented. The results are analyzed by a person who does not know whether the participant was in the treatment condition or the control condition. If the outcome of an efficacy study of an intervention shows that the intervention group did better than the control group over a number of trials, the procedure can be identified as empirically-supported therapy. Whether the procedure investigated in the efficacy study actually works in practice, which includes the vagaries of the intervention, has to be tested in an effectiveness study (i.e., in the field).

Reference

Duggan AK, McFarlane E, Fuddy L, Burrell L, Higman SM, Windham AK, & Sia C (2004) Randomized trial of a statewide home visiting program: Impact in preventing child abuse and neglect. *Child Abuse & Neglect* ,28:597–622.

Reading the Limitations of a Research Study

There are many potential sources of error in the design, execution, analysis, and reporting of research results. Some of these errors have been previously discussed and include confounding, bias, sampling, and the role of mediators and moderators. Another limitation in published research is writing that misleads or has the potential to produce misunderstanding or overgeneralization of research results. Most journals require authors to include as a part of their paper a statement of the limitations of their research.

Better understanding of statistical concepts can help the reader interpret the results of a research study in a more complete fashion that merely accepting the author’s conclusions. Such understanding will lead to better presentations of data to other interested parties and better prevention and treatment practices when these based on research results.

In the article entitled “Extent, nature, consequences of rape victimization: Findings from the National Violence Against Women Survey” (Tjaden & Thoennes, 2006), the authors include such a section on limitations. The limitations they note are the following.

1. The small number of women (24) and men (8) in their survey who had been raped in the past 12 months in their representative sample. The authors advised interpreting the results with caution.

2. The survey did not include rapes of children, adolescents, those living in institutions, and the homeless. The authors advised that the study under-

estimates the prevalence of rape by not sampling populations where this may occur more frequently.

3. Since the study was conducted by telephone those persons without a telephone were not included. With the changes in communication technology (such as computers, cell phones, and other devices) future survey research may become much more complex and introduce known and unknown biases. For example, cell phone numbers are not currently published, but there are private agencies that provide lists of cell phone numbers. It is not clear to what degree such lists include typically underrepresented portions of the population.

4. The impact of race and ethnicity in surveys is a difficult issue to understand. Some groups such as Native Americans and Asians have such small populations in the U.S. that getting an adequate sample is difficult (if not impossible) for small surveys. This was the case in this study of Asian/Pacific Islanders. Hence, results for a group with a small number of respondents should be viewed with caution. The reader should also be careful about interpreting results from a survey of low frequency events and selected populations unless the survey is large and the mechanism for ensuring representation is carefully explained.

5. Finally, to have a good understanding of survey results, one needs to know (a) exactly what was the question, and (b) how are events defined. In the Tjaden and Thoennes (2006) paper, to their credit they report survey definitions and questions. However, one of our editors noted the rape statistics that reported were higher than published elsewhere. If one reads only the introduction or summary of the findings, the reader would miss the definition of rape (for this survey) as being *either attempted or completed rape and the use of or the threat of force*.

6. It is important to always read the author's description and consider the limitations of any study. There are always limitations and this is one reason why repeating studies with different methodologies and in different populations is so important.

Reference

Tjaden P & Thoennes N. *Extent, nature, consequences of rape victimization: Findings from the National Violence Against Women Survey*. National Institute of Justice Special Report No. 210346.

Section 6

Special Topics

In this section we present topics that do not fall into one of the above categories or they involve more than one. The first topic is a discussion of the differences between Army and civilian domestic violence rates. This topic comes up periodically, particularly when the FAP draws media attention. This article provides a summary of relevant information for FAP personnel to use, when necessary, to improve public understanding of the research and issues involved. Since needs assessment is a requirement of the Army Community Service, we consider some ways of conducting them. We focus particularly on the issues of sampling and the populations available for sampling. The next article is oriented toward understanding evidence-based research and how such research applies to clinical practice. Efficacy (does the treatment work in a controlled setting?) and *effectiveness* (does the treatment work in actual clinical practice?) have been discussed in many articles in this series. In this article we review a paper that argues for fidelity in following treatments that are evidence-based. In our final article, we review the responsibilities of the Army Family Advocacy Research Subcommittee (FARS) and the requirements for submitting proposals and protocols to it for review. This material is intended to help investigators who plan to do research. It can also be helpful for managers who supervise individuals who are considering research topics.

Comparison of Army and Civilian Spouse Abuse Data

There have been a number of recent citations in the news of comparisons between the rates of domestic abuse in the military and the civilian community. Some stories indicated that the military rates are higher by a certain percentage while others just say that they are greater. For example, the Christian Science Monitor on 5 August 2002 reported that soldiers are about twice as likely as civilians to turn violent at home. Currently, there is only one study that compared the rates of self-reported spousal aggression in military and civilian populations (Heyman & Neidig, 1999). This study is often referred to as the *comparability study* since its purpose was to compare the rates of domestic violence in the Army to the U.S. national rates.

Heyman compared data collected via a paper-and-pencil survey administered by Peter Neidig during 1990–1994 at 38 Army installations to that of the 1985 National Family Violence Survey (Straus & Gelles, 1990). Heyman found no statistically significant difference in the adjusted rate of moderate spousal aggression (about 10%) between the Army and the civilian data. However, for severe spousal aggression, the adjusted rate for the Army was 2.5% and was 0.7% for the civilian data. These differences were probably due to race (more minorities) and the young age of the Army population and not to abuse propensity.

There were substantial methodological differences in the two studies. The data were collected over different periods of time. They used different sampling methods and data collection techniques, and the demography of the samples was very different. Therefore, the ability to compare across the two studies was limited and was statistical and not direct. Separate analyses were done by sex and were further stratified by age group and race.

The civilian sample included only married employed respondents under the age of 65 from both groups. All of the Army population sampled was on active duty. The final comparison was based on an adjustment of the Army sample to the 1990 U.S. Census. (This was required in order to be able to say what the rate of violence would be in the Army if it had the demographic structure of the U.S. Census in 1990). Straus and Gelles used telephone sampling and had a low percentage of some groups of respondents. A low response rate was observed with young minorities, a particularly important point considering the high rate of domestic violence in that population. Despite the rigor in matching the samples and in weighting them to the census, statistical comparison could not replace missing data. Therefore, while this was the only possible comparison, it had significant limitations. We emphasize this not to criticize Dr. Heyman, but to indicate that this was the only possible method of achieving any comparison because of the cited differences. In addition, the Army data are more than a decade old and the civilian data used in this study is more than 15 years old. Finally, these are self-report data and were not related to reported cases of domestic violence.

One of the reasons for comparing military and civilian data using these two very different samples was because there is no centralized national database of actual reported domestic violence cases by which one might make comparisons between military and civilian populations.

The studies cited use prevalence data. These data were collected from a sample survey using the Modified Conflict Tactics Scale (CTS), which asks the respondent to describing how that couple might have resolved conflict during the past year. Prevalence data provides an estimate of the aggression in the population and is different from a database of actual cases. Our conclusion is that this study was a good attempt at the time to compare military and civilian spousal aggression, but it is dated and has inherent limitations. Given the methodological and demographic differences, it is uncertain whether these data are representative of either the military or civilian populations of today.

References

Heyman RE & Neidig PH. (1999). A comparison of spousal aggression prevalence rates in U.S. Army and civilian representative samples. *Journal of Consulting and Clinical Psychology, 67*:239–242.

Straus MA & Gelles RJ. (1990). *Physical violence in American families: Risk factors and adaptations to violence in 8,145 families*. New Brunswick, NJ: Transaction.

Army Community Service Needs Assessment

In general, a needs assessment is a procedure to match resources to the needs of the population being served. Since ACS is in the human services field, a needs assessment of all of its programs could become extremely complex if reasonable limits are not imposed on its goals and objectives.

ACS has many programs: the Family Advocacy Program (FAP), Mobilization and Deployment Support (MDS), Army Family Team Building (AFTB), Relocation Readiness Service, Financial Readiness, Employment Readiness, the Exceptional Family Member Program (EFMP), Volunteers, Army Family Action Plan (AFAP), and many more.

The FAP prevents and treats child and spouse abuse by providing a variety of services to strengthen families. MDS provides support services during all phases of deployment to all eligible family members. AFTB is an educational program to promote personal and family readiness. Financial Readiness teaches soldiers self-sufficiency in their financial affairs. Employment Readiness helps soldiers who are leaving the Army who may have employment problems associated with their relocation and transition. The EFMP supports families with special needs. The Volunteer program augments staffs and expands program capabilities through individual donations of time and work. The AFAP (through efforts at the installation level) identifies services and initiatives critical to improving standards of living in the Army.

Why should you conduct a needs assessment? Needs are always greater than resources. One reason for the existence of governments and bureaucracies to allocate resources to community needs. Another reason is to monitor the processes of change within communities served by Army programs. The Army can change in a variety of important ways depending on such factors as the economy, changes in American society, changes in laws and regulations, education, and demographics (age, race, and sex). Therefore, a program that was tailored for a group five years ago may not currently serve that same population. In a large organization like the Army, people who provide funding (Congress, DoD, the Army) are a long way from those who implement and participate in programs. Therefore, connecting these two mechanisms in a timely fashion is often a challenge.

What are the methods one might use in a needs assessment? The primary methods are surveys (via paper and pencil, or electronic media), interviews (in person or telephone, individual or group), and the collection of data from other sources that relate to the program. Data should be gathered through a process that allows every person the same opportunity to be selected. This is extremely difficult to do in practice because of the difficulty in (1) identifying members of the population, (2) contacting them, and (3) obtaining valid data. Rather than having a random sample from which one might draw conclusions about the population (within reasonable limits), you may have to use a convenience sample. Regardless of the approach, a needs assessment must be done systematically.

What are the measures that you use? Can you ask just a few questions that will give you meaningful data or do you need to ask many questions? Obviously, there is a tradeoff between collecting as much data as the researcher needs and the time participants want to engage in the research. Is there a standard that you can use to measure effectiveness? Is a program that reaches 10% of your available population with a 90% level of satisfaction any better or worse than a program that reaches 50% of the population with a 50% satisfaction level?

Surveys and interviews could be done with many different groups of people to gather data for the needs assessment. In the military community, one could talk to commanders, program managers, participants, and nonparticipants. Outside the military, one could talk to providers of civilian programs, which are frequented by military families and other collateral resources. In addition to the aforementioned question, the design of a needs assessment raises several other questions. What problems/issues occur when trying to match resources to needs? What measures will you use?

- Do you only focus on persons who have participated in programs or the entire post?
- Do persons with the greatest need participate in the program?

- Can the program meet participants' needs without stigmatizing them?
- Is the program available at a time when individuals can participate?
- Are efforts made to periodically or continuously recruit individuals who might benefit from the program?

It is likely that the needs assessment will be a balance of the responses of individual program participants, program managers, community leaders, and data from existing records. ACS directors will then be able to collect and interpret assessment data in a way that will benefit their installation.

Efficacy and Effectiveness Research and Its Impact on Couple and Family Therapy

In this article, we address the impact of research on the clinical practice of couple and family therapy. The Army Family Advocacy Program (FAP) has emphasized the importance of outcome research, particularly with regard to whether various treatment modalities for family violence actually work. One of the most widely used treatments for family violence in the Army is couple and family therapy (CFT). In this article, we review a paper by Pinsof and Wynne (2000) on their conceptualization of the relationship between research and clinical practice. Their thesis is that CFT research, as currently conducted, has had little impact on real life clinical practice because couple and family therapists do not consistently adhere to the rigid methodological demands called for in treatment-focused research. Generally, researchers are trained to use pre-set criteria, treatment manuals, and their work under controlled conditions with clients is usually monitored. On the other hand, clinicians are inclined to use eclectic, integrative, or multimodal methods of treatment based upon their perception of their clients' needs.

Often, there is a sense among therapists of needing a treatment model that works based upon clinical trials and scientific evidence. The question of whether the therapy works is called *efficacy* research. Efficacy research involves six primary elements: (1) a clinical laboratory setting; (2) a focus on a definable disorder or condition; (3) the presence of a treatment group and a control group; (4) the random assignment of clients to one of these two groups; (5) manualized treatment (i.e., therapy that is conducted using a standardized technique that is described in a manual on which the therapist is trained to a criterion) that is monitored during the therapy; and (6) pre-post therapy measures of some aspect of client functioning such as feelings and behaviors. Basically, efficacy research asks whether treatment is better than no treatment at all or whether one method of treatment is better than another.

Almost all the reviews with which we are familiar indicate that psychotherapy is generally thought to be better than no therapy. Some therapies have been shown to be better than others, and some studies indicate that a combination of therapies (e.g., medication and cognitive-behavior therapy for some forms of depression) are better than either type of singly applied treatment.

Effectiveness research follows the establishment of efficacy. It attempts to determine if the treatment that was found to be effective in a laboratory setting actually works in real-life practice. Such research would still be somewhat removed from practice because of its reliance upon a uniform concept of treatment, i.e., that the therapy was uniformly applied by all therapists and in all cases. Pinsof and Wynne support a definition of effectiveness research that differs in two ways from that just described: it does not have to invariably follow efficacy research (because it may be impossible or impractical to conduct such treatment with certain populations) and it does not necessarily have to be based on manualized treatments. This second condition makes this form of effectiveness research less radically different from the way clinicians actually practice.

Pinsof and Wynne believe that what is needed in CFT research is a study of three elements:

(1) how family change occurs naturalistically,
(2) how families change in therapy, and
(3) how to develop strategies to identify therapist interventions and in-therapy experiences that can be linked to client change.

How is the information in this article relevant to the Army FAP? The article questions the argument for standardized training and practice interventions and supports therapists' use of practice experience. This experience would be put to use in being observant on what actions make a difference to couples and families in and outside of therapy and what cues the therapist uses to guide the therapeutic process. Pinsof and Wynne describe therapy as essentially an ideographic process – one that is organized in regard to the individual and is based on a continual change of course in response to the cues provided by the client. Therapy is also seen as an educational activity, in which the therapist encourages clients to think, feel, or act differently in regard to themselves and others.

Pinsof and Wynne present a research model that they believe is clinically relevant and can change and inform treatment. As an alternative to treatment-focused efficacy and effectiveness research, they propose the use of a client-focused learning process research model. They believe that this model will generate information to assist therapists in determining and influencing the progress of cases in the change process.

There are many possible approaches to conceptualizing and designing research that will contribute to your understanding of how to help FAP clients. We encourage you to consider the work of Pinsof and Wynne for an explanation of the relationship between research and real life clinical practice with FAP clients.

Reference

Pinsof WM & Wynn LC (2000). Toward progress research: Closing the gap between family therapy practice and research. *Journal of Marital and Family Therapy, 26*, 108.

Submitting and Evaluating Research Plans

The Army has an organization that reviews and approves research plans for family advocacy-related research that is performed by Army investigators or by others who want to conduct research using Army populations. This organization is the Family Advocacy Research Subcommittee (FARS). The FARS was organized under the auspices or the Department of the Army (DA) Family Advocacy Committee to review, coordinate, and recommend the approval and dissemination of family violence research (AR 608-18). FARS has the responsibility for all research activities related to the Army (FAP) world wide and use of the Army Central Registry. Additionally, the FARS Standing Operations Procedures (SOP), DA regulations and numerous supplements exist at the major command and installation level to provide guidance on conducting FAP research. In some cases, the definition of research itself may be at issue. There are instances in which program reviews and other types of clinical and administrative reviews may be exempted from review as research. The FARS will not routinely review recurring installation program evaluation and analyses such as customer satisfaction surveys, internal reviews, quality assurance assessments, management information system analyses, and annual reports. The FARS reviews all FAP studies intended for publication in a scientific journal or book.

The FARS SOP has two sets of guidelines: one for proposals and one for protocols. The proposal format exists so that an investigator can submit a skeleton plan (an idea) to see if the FARS is interested in the focus of the research and wants the investigator to further develop the idea. If an investigator receives encouragement, then the next step is to prepare a protocol. A protocol is a research plan that includes a literature review, a statement of hypotheses, a research design, and statements on what the research is likely to achieve. The literature review tells what is known about a specific problem (say the effects on children of witnessing spouse abuse). Very frequently, but not always, a research hypothesis is examined statistically.

Following the literature review and statement of hypotheses, the investigator tells the reader how the study will proceed. This will include statements on who the subjects are, what sort of data will be collected, how it will be collected (e.g., mailed questionnaires or interviews of patients), and how it will be analyzed. All of these involve extensive preparation. Some studies may also require a preliminary investigation that is conducted in a pilot study which may or may not be possible depending on the installation's policies. If the analysis includes statistical tests, how many subjects will be needed? This can be very tricky, particularly when you begin to break the analysis down by groups such as by males and females, type of intervention, treatment, or prevention program, active duty and civilian status, and other classifications. The numbers of required subjects can add up quickly. In addition to these considerations, the data collection instruments require close examination. Such issues as reliability and validity are the primary measurement concerns, assuming that people volunteer for the study. If an investigator expects to receive funding for the research, a budget must be submitted so that adequate funds will be available to accomplish the mission. Finally, there must be some idea of what the research will contribute to the Army FAP.

When reviewed by the FARS, proposals/protocols are evaluated relative to the:

- Compliance with the administrative criteria of the FARS
- Scientific and programmatic relevancy and quality
- Experience of the investigator
- Soundness of the literature review
- Consideration of human use issues and other specific administrative issues, if necessary
- Research design including the plan for data collection and analysis
- Reasonableness of the budget
- Contribution of the research to the FAP

Investigators may be invited to a meeting of the FARS to discuss or clarify their research interest, methodology, and the relationship of their literature review to their inquiry. The FARS permits revisions of proposals and protocols. The FARS procedures are detailed in a SOP, which is available from the FARS.

All proposals and protocols from Department of the Army personnel for studies and research projects involving human participants involved in family advocacy issues; e.g., physical, emotional, and psychological spouse/child abuse must be routed through Headquarters, U.S. Army Medical Command, ATTN: MCHO-CL-H, 2050 Worth Road, Suite 10, Fort Sam Houston, TX 78234-6010, to Commander, Family Morale Welfare and Recreation Command, ATTN: Family Advocacy Program Manager, 4700 King Street, Alexandria, VA 22302-4418. Prior to final approval of a protocol by the FARS, Army personnel must include written evidence of review and approval from their local Institutional Review Board (at the relevant US Army Medical Center) and the U.S. Army Medical Department Center and School, Clinical Investigative Regulatory Office. All others must submit proposals directly to the Army Family Advocacy Program Manager. Non-Army personnel must likewise submit evidence that they have complied with the rules of their Institutional Review Board.

Glossary/Definitions

Bias– Bias in statistics refers to a systematic error in obtaining information such as in data collection. There are many types of bias (recall, selection, observational, interviewer, misclassification, sampling). Some are known ahead of time and can be addressed statistically while others are unknown.

Central tendency – A general term for specific properties of a distribution of observations. The most common measures of central tendency are the mean, median, and mode.

Chi-square test – A test of statistical significance used for categorical data, which may be frequencies, percentages, or proportions.

Cluster analysis – A type of correlational procedure used in the analysis of large amounts of data. Observations (data) are classified into groups based on shared properties such as how they relate to other statistically (as in simple correlations). The results of the analysis provide categories that allow the reader to more easily understand the characteristics of the data.

Confidence interval (confidence limit) – One of the most basic statistical questions is whether there is a statistically significant difference between two sample means. In order to test this, one has to know how much variation is in the measures. Confidence intervals tell you how what are the expected limits of your sample mean based on a selected level of probability. A confidence interval is constructed based on the amount of variation in the sample and the level of confidence the investigator is willing to accept. The most frequently used confidence interval is at the 95% level.

Confounding – A type of systematic bias that occurs when the effect (result) of interest is mixed with the effect of another factor. Confounding can lead to biased estimates of an effect depending on the relation of the confounder to the exposure and the result.

Degrees of freedom – The degrees of freedom are part of the calculation of statistical significance, which tests the probability of a hypothesis being rejected or not rejected. In one example of how degrees of freedom are calculated, the sum of the deviations from the arithmetic mean must add to zero. For example, take the mean of 5 numbers: 2, 4, 7, 8, and 9. The mean is 6. This value (6, the sample mean) is fixed as your estimate of the population mean. The deviations from the sample mean are -4, -2, +1, +2, and +3. The sum is zero. If you keep your sample mean fixed at 6, you can change four of the estimates, but the 5th cannot be changed or the sample mean will not be the same. For example, estimates can be changed to 1, 3, 5, and 6. In order to keep the sample mean at 6, the final number has to be 15. In order for the sum of deviations to add to zero, they must be –5, –3, –1, 0, and +9. Thus, the number of degrees of freedom is N-1. The minus one denotes the observation that cannot vary. Degrees of freedom can be illustrated for different types of calculations and tests, but the concept of how much variation can occur is the same regardless of the test of significance.

Effect size – The difference between the results obtained due to an intervention and the result without the intervention. The concept of effect size can be statistically complex. In its simplest form, it is meant to convey a sense of how large is a result (an perhaps how meaningful) in contrast to statistical significance. A result may be highly significant statistically (i.e., have a very small p-value), but trivial in terms of how big an effect is observed. [See statistical significance]

Effectiveness – Effectiveness research aims to determine if a treatment that has been found to be effective under controlled conditions will work in actual practice in another location such as someone's practice. (Also see efficacy.)

Efficacy – In the context of psycho-social interventions based on evidence, the first research strategy is to determine whether the treatment works. Efficacy research is conducted under controlled conditions to see if the intervention works under the best possible circumstances before it is moved out into practice (see effectiveness).

Epidemiology – Branch of medical science devoted to understanding the causes and effects of diseases. In addition, while not limited to epidemiology, it can lay claim to specific language and statistical concepts to communicate its methods and results. Among these are exposure, classification, association, bias, confounding, and stratification, to name a few.

Exposed-NonExposed – This is the most basic classification in the design of epidemiological research in which one group (the exposed group) has a particular event or characteristic of interest (e.g., substance abuse) and the other group (the non-exposed group) is free from that event.

False negative – A result on a test (e.g., a screening test for a disease or condition) in which the test result is negative (test negative), but the individual has the disease or condition. Thus, while the test is positive, the result is false because the test failed to correctly identify the person correctly.

False positive – A result on a test (e.g., a screening test for a disease or condition) in which the test result is positive (test positive), but the individual does not have the disease or condition. Thus, while the test result is positive, its results are false because the test fails to correctly reject the person.

Frequency – The number of events or observations in a distribution.

Gold standard – The term commonly used in research and in clinical care (such as in screening for disease) and in the media, to denote the highest possible level of value when referring to a standard. It is usually used for the purpose of comparison. If an event meets a gold standard, it is considered to have met the gold standard.

Hypothesis testing (H_O and H_1) – The purpose of an experiment is to reject the null hypothesis. H_O is the notation for the null hypothesis and H_1 is the alternative hypothesis. H_O can never be proven, only accepted or rejected. If the test of significance meets the probability level set by the investigator (usually no more than a probability of 0.05 that the null hypothesis it is true) then H_O is rejected (see Type I and Type II errors). H_1 can be vague or specific. In general, H_1 only states that there is a difference between two means. Further specificity can be obtained by hypothesizing the direction of the difference [see one and two-tailed tests], but typically the investigator will test both possibilities and not specify the direction of the alternative hypothesis.

Incidence – The most basic measure used in epidemiology. Incidence refers to the number of new cases per unit time. It differs from prevalence in that for an incidence figure to exist, time must pass. For example, we may speak of the number of new cases of HIV per year.

Incidence rate – More informative than simply incidence is the incidence rate, which is the number of new cases per unit of time and per unit of population. There are many complex approaches to measuring the incidence rate, but this is the basic concept.

Mean – There are different types of means, but the mean usually used in statistics is the arithmetic mean, the sum of the observations divided by the total number of observations.

Median – The point on a normal distribution at which half the observations fall below and half fall above.

Mediator – A factor on the pathway between the independent variable (the factor you are interested in studying) and the dependent variable (the outcome). A mediating variable must demonstrate a significant degree of relationship between the independent and the dependent variable. If no relationship exists, then the hypothesized mediator does not lie on the causal path and hence cannot be a mediator.

Mode – The most frequent observation in a distribution of numbers.

Moderator – A characteristic that exists only in certain subgroups that have a relation to the dependent variable (the outcome). These characteristics make the relationships between independent and dependent variables stronger or weaker. In contract to mediating variables, a moderating variable should have little or no statistical relationship to either the independent or the dependent variable.

Negative predictive value – In screening test results, one can predict the likelihood that a person without the outcome for which the test is given actually does not have that outcome. It is expressed as a probability and is calculated by taking the number of true negatives divided by the sum of true negatives and false negatives.

Odds ratio – A type of rate ratio that is a measure of effect size obtained in case-control studies of disease incidence. It is obtained by comparing the odds of an outcome of interest occurring in two different groups. It is calculated as the ratio of cases to controls among the exposed subjects divided by the ratio of cases to controls in an unexposed group of subjects. It is sometimes called a cross-product ratio.

P-value – A value selected by the investigator when setting up a statistical test. The p-value is the probability of rejecting the null hypothesis (Ho) when it is true.

Percentage – The parts out of 100. It can be expressed as a fraction or a ratio where it is understood that 100 is the denominator. For example, .50 is understood as 50%.

Positive predictive value – In screening test results, one can predict the likelihood that a person with the outcome for which the test is given actually has that outcome. It is expressed as a probability and is calculated by taking the number of true positive divided by the sum of true positives and false positives.

Prevalence – The amount of occurrence of an event in a population. It can be expressed in a variety of ways such as a proportion, percentage, or rate. Regardless, it is a measure of how much of some event or characteristic is present at a given time.

Prevalence rate – The amount of an event per unit of population.

Proportion – The relation of a part to a whole. Proportions are often expressed as ratios. For example, quantities are in proportion in the following relationship: 1/3 = 2/6. It is not the same as a percentage.

Random – In statistics, in drawing a sample each member of a population has an equal chance of being selected. Randomization is a procedure used to attempt to reduce bias.

Rate ratio – One of many measures of a relative effect. It is the strength of an association where the numerator is the quantity of interest and the denominator is the reference or baseline. It is expressed as the ratio of two rates, the numerator in relation to the denominator minus one. The integer one (the minus one) is subtracted from the rate ratio thus leaving it as the excess effect above zero. For example, when the two ratios are equal, the rate ratio is one. When the integer one is subtracted from the unity measure of the rate ratio, the result is zero, which it should be if there is no effect. The rate ratio is often termed the relative risk or the relative rate.

Ratio – The relation of two quantities such as in a fraction.

Receiver Operating Characteristic (ROC) – Used in the evaluation of cutoff points for screening tests, the ROC is a plot of test outcome characteristics such as the sensitivity compared to the specificity or the true positive rate compared to the true negative rate. The ROC is a visual method of selecting a screening test model (such as by using a cutoff point) to provide the optimal result given the screening test's ability to discriminate true positives from true negatives.

Relative risk – See rate ratio.

Reliability – A measure used in test development to characterize the consistency of measures of an outcome. There are many types of reliability such as test-retest (the same test is given again under the same conditions at a later point), split-halves (a test is divided into halves and compared), and alternate forms (two supposedly equal forms of a test are compared) and internal consistency (the degree to which items on a test or survey agree with each other in assessing the concept that is measured).

Risk assessment – A term whose meaning varies depending on the subject matter. In the health field it refers to the process of examining and measuring the harmful effects of a characteristic of a population or of an individual. Measurement can be quantitative or qualitative.

Sampling – A procedure used in statistical estimation to collect data on a characteristic from a segment of a population in which it is not possible or feasible to measure all individuals. There are many types of sampling, but all aim to estimate a property of the entire distribution of observations of interest.

Screening – In the health field, screening tests are given to determine the presence or absence of a condition such as cancer. (Often, a screening test is just the first step in determining if an individual has an outcome. A positive screening test can be followed up with a more specific test such as a biopsy.) Because tests are not infallible, statistical procedures are utilized in developing screening tests to determine the probability of positive and negative outcomes. (See also sensitivity, specificity, positive and negative predictive values, false negative and false positives, and gold standard.)

Sensitivity – The ability of a test to identify if a person has the outcome condition. Sensitivity is calculated by taking the number of true positives (people with the outcome) and dividing by the sum of true positives plus false negatives (people who have the outcome, but do not test positive).

Significance (significance level) – The level of probability in a statistical test that the difference between the two means is not due to chance is called the level of statistical significance. It is usually accepted at 5% or less. Statistical significance is determined by several factors: the actual difference between the means (greater is better), the amount of variability of the measures of the two means (less is better), and the number of observations (the more observations collected in a sample, the more likely it is that the shape of the true distribution is approximated).

Specificity – The ability of a test to identify if a person does not have the outcome condition. It is calculated by taking the number of true negatives (people without the outcome) and dividing by the sum of true negatives plus false positives (people who test positive, but do not have the outcome).

Standard deviation – A measure of variability in a set of observations that is used in testing the significance of differences between means. It is calculated by taking the square root of the variance.

Stratification – In analyzing data from a large population, variability can be reduced when sub-elements of the population are considered separately. Examples of strata are age, race, and gender.

T-test – A statistical procedure that determines if the difference between two means is likely to be different from chance.

Type I error – A type of error that can occur in testing the significance of differences between means. A Type I error occurs when you reject the null hypothesis (H_O) when it is true. This amount of the Type I error is stated as the significance level you are willing to accept. The level of 5% ($p<.05$) is the usual standard for statistical significance. This level of 5% means that you are willing to accept the risk due to chance that you are wrong in rejecting the null hypothesis 5% of the time, or 5 chances out of 100.

Type II error – A type of error that can occur in testing the significance of differences between means. Type II error occurs when you fail to reject the null hypothesis (H_O), failing to find a difference when one is actually there, there actually is a difference, but you do not detect the difference due to chance.

Validity – Validity is a concept in test development that indicates the degree to which you are measuring the concept that you are attempting to measure.

There are several types of validity. The most common are concurrent, convergent, discriminative, predictive, criterion, and face validity.

Variance – A measure of the dispersion of a distribution of observations. It and the standard deviation (the square root of the variance) are used in many statistical tests to measure variability of distributions and hence test the significance of differences. It is calculated by taking the average squared difference between each observation and the expected value. (The expected value is calculated differently in various statistical tests, such as chi-square, but it represents the overall value an observation would have if the experiment were conducted many times.)

Useful References

Rothman KJ. (1986). *Modern Epidemiology*. Boston: Little, Brown, and Company.

Rothman KJ & Greenlander S. (1998). *Modern Epidemiology*. Philadelphia: Lippincott-Raven Publishers.

Everitt BS. (2006). *The Cambridge Dictionary of Statistics.* London: Cambridge University Press.