

This section includes shorter (e.g., 10–15 double-spaced manuscript pages or less) papers describing methods and techniques that can improve evaluation practice. Method notes may include reports of new evaluation tools, products, or services that are useful for practicing evaluators. Alternatively, they may describe new uses of existing tools. Also appropriate for this section are user-friendly guidelines for the proper use of conventional tools and methods, particularly for those that are commonly misused in practice.

Making Statistics More Meaningful for Policy Research and Program Evaluation

HENRY MAY

Available online 29 September 2004

ABSTRACT

This article focuses on the use of statistics in policy and evaluation research and the need to present statistical information in a form that is meaningful to mixed audiences. Three guidelines for formulating and presenting meaningful statistics are outlined. Understandability ensures that knowledge of statistical methods is not required for comprehending the information presented. Interpretability ensures that statistical information can be explained using familiar, non-abstract units. Comparability ensures that the magnitudes of different estimates can be directly compared within and across studies. Examples for improving popular effect size estimates from linear and non-linear models are included, and a general approach to presenting statistical information meaningfully for consumers of policy and evaluation research is explained.

Henry May • Consortium for Policy Research in Education, University of Pennsylvania, 3440 Market Street, Suite-560, Philadelphia, PA 19104, USA; Tel: (1) 215-573-0700x236.; Email: hmay@gse.upenn.edu (H. MAY)

American Journal of Evaluation, Vol. 25, No. 4, 2004, pp. 525–540. All rights of reproduction in any form reserved.
ISSN: 1098-2140 © 2004 American Evaluation Association. Published by Elsevier Inc. All rights reserved.

INTRODUCTION

The best policy and evaluation research exhibits a combination of expert analytic skills and well-developed communication skills. An individual researcher might be good in both, but if the study involves modern statistical methods, the link between analysis and communication of results often is most difficult to build. Explaining the results of statistical models in a manner that is understandable and acceptable to multiple audiences can be a difficult task. Details about the methods are important to fellow researchers, while at the same time, policy makers and the general public may be more likely to appreciate the results when findings are explained in ways that are understandable to the layperson. In a conflicting trend, an increasing number of researchers are using confusingly complex statistical models to test research hypotheses. Even so, results ultimately must be interpretable at a level that requires no knowledge of the statistical techniques used. Clearly, the method of analysis is only a means to an end. Statistics are simply tools for addressing a research problem or question, and policy and evaluation research questions are customarily phrased in plain language—shouldn't our answers be equally free of analytic jargon?

One might consider each research endeavor as a continuum or process that has a question and conclusion, a beginning and end; and that the product of this process is most useful when the continuum comes full circle, and the beginning and end are linked. Then the conclusion is on the same level, and just as understandable as the question. For simplicity's sake, imagine such research as a process consisting of three stages. In the first step, a research question is identified and stated clearly. What is the nature and severity of problem A? What are the impacts of program B? How strong are the relationships between X , Y , and Z and an outcome of interest? Step two involves investigation and analysis to develop answers to the research questions. If our research design has quantitative aspects, this is where statistics enter the scene. Analyses of varying complexity are performed, and parameter estimates are produced. It is what comes next, the transition to the third stage of reporting results, that is critical to coming full circle in the research process. If this transition is done abruptly, and the results are presented in a raw form, then the research article or evaluation report is really useful only to those who are familiar with the statistical method used and its output. This clearly should be avoided in the contexts of policy research and program evaluation. Utility of the results of such research should be maximized. Therefore, attention and care should be afforded to this transition process of converting the raw results of statistical analyses into a more intuitive form. Otherwise, the only people who can have a good understanding of the effects of a policy or program are those who are familiar with the statistical terminology used to describe the results. My experience tells me that this happens all too often in policy and evaluation research.

The purpose of this article is to illustrate several techniques for making statistics more intuitive and meaningful in the context of policy and evaluation research and to identify general themes common among these methods that can serve as guidelines for other analyses not mentioned here. This paper does not include an exhaustive review of the literature on the topic of communicating quantitative findings. It simply describes and illustrates several useful practices and principles, while recognizing a few promising methods developed by others. Furthermore, the examples presented in this article focus primarily on statistical analyses involving regression

models or ANOVA, which are very common methods in evaluation research. However, the general ideas presented here surely extend to any statistical analysis.

GUIDELINES FOR MEANINGFULNESS

It is helpful to identify key characteristics of meaningful statistics that may serve as guidelines or goals for directing creative and effective presentation of results. Emphasis on certain goals over others would depend upon the context of the particular research topic and the audience for which it is intended. Particular approaches are very helpful when the audience is largely non-technical, while other approaches build upon an audience's preexisting knowledge of statistics or probability with varying degrees of complexity. The guidelines outlined here are designed to be general enough to encompass most applications; however, the ability to satisfy each guideline will depend upon the form of the statistical results and the creativity of the researcher.

Guideline 1: Understandability

The first guideline essential to making statistics meaningful is to improve the understandability of results. Understandability is crucial for maximizing the utility of policy research findings. *The results should be reported in a form that is easily understood by most people by making minimal assumptions about the statistical knowledge of the audience and avoiding statistical jargon.* If findings are riddled with statistical terminology like "regression coefficient", "standardized effect size estimate", or "variance component", the researcher risks alienating readers who might not have the necessary statistical knowledge to interpret the findings. Researchers in the policy and evaluation arena should recognize that public policy is heavily influenced by public knowledge and opinion. If we ignore the general public as potential consumers of research, we are drastically limiting the ability of our research to inform public policy and promote effective programs. Furthermore, it is crucial to recognize that making statistics understandable does not imply "dumbing-down" the level of information. It simply means that statistical jargon is reduced or eliminated, and inferences are explained in plain English (or French, Chinese, Spanish, etc.). This allows readers to understand the findings directly through clear, intuitive, and straightforward expression.

Guideline 2: Interpretability

Another important guideline for making statistics meaningful is to enhance their interpretability. *A statistic is interpretable when the metric or unit of measure that it is based upon is familiar or easily explained.* Ideally, interpretability follows the same model as understandability. Can the units of the statistics being reported be explained in plain language and do they make only minimal assumptions about the statistical knowledge of the audience? For instance, a researcher evaluating the impact of a new curricular program could report how much children's grades or test scores increased in terms of actual points on the state test (which is probably a familiar metric for most parents and policy-makers in the state) while also reporting the standardized effect size (which will be of interest mainly to fellow researchers). The interpretability of a standard error for an impact estimate is improved dramatically by reporting confidence intervals in the original metric of the outcome, "The average improvement in test scores of students participating in the program ranges from 30 to 80 points." Results from nonlinear regression models (e.g., predicting proficiency levels

on a test) are often reported in terms of odds, a statistical metric that is somewhat understandable and interpretable. However, it is relatively easy to produce predicted probabilities from non-linear models, and the direct comparison of probabilities or proportions is easier to explain than odds or odds ratios. By identifying interpretable metrics, researchers reduce the chances of alienating important audiences who are unfamiliar or uncomfortable with statistical terminology.

Guideline 3: Comparability

A third guideline for making statistics meaningful, which may be less crucial than the first two, is comparability. This simply means that the reported sizes of statistics that might be compared can be compared directly, without any further manipulation. These comparisons might be made across different factors in a single study, to analyses involving different dependent variables, or to effects from other studies entirely. Comparability is most often a concern when reporting coefficients from a linear or non-linear regression. Standardized regression coefficients are commonly used to achieve comparability; unfortunately, however, they require prior knowledge of statistics and their interpretation is not intuitively obvious. With creativity, the policy researcher can make a few small changes to the presentation to get comparability (an example is presented later), while limiting the use of statistical jargon and making minimal assumptions about the audience's statistical knowledge.

The key to meeting all three of these guidelines involves rethinking the way we present statistics, and presenting results in ways that capitalize on more commonly understood statistical metrics. Current practice often entails the presentation of traditional statistics using visual means such as graphs or figures. Although these visual tools can be incredibly effective in relaying major points, they do not necessarily reduce reliance on statistical jargon or unfamiliar metrics. This makes it difficult for anyone to use the information presented if the figure is not in front of them. The figure becomes much more powerful when those who have seen it can explain its message to others using familiar, straightforward language. However, the task of reporting statistical results in straightforward language is not always simple, and the goal of making every aspect of results salient to all audiences may be unrealistic. It is important to consider the target audience, the intended message, the potential uses of results, and other contextual factors when deciding what to report and how to report it.

THE BENEFITS OF MULTIPLICITY IN PUBLISHING FOR MULTIPLE AUDIENCES

Consumers of evaluation and policy research vary widely in terms of their familiarity with statistical methods. If researchers were interested only in publishing results so that other researchers could read them, then making statistics meaningful would be much less of an issue. Yet, reporting results using only traditional statistical language would leave non-technical audiences with a very limited amount of information, sometimes devoid of useful quantitative estimates. This is unacceptable when it is relatively easy to report quantitative results in a more meaningful way.

Research evaluating the effectiveness of programs and policies in education may have the widest potential audience, ranging from the very technical education research community, to less technical policymakers, to teachers and administrators, and finally to the largely non-technical communities of parents and students. In order to maximize the utility of their findings,

education researchers must consider each of these stakeholders and provide each with useful and meaningful information. Sometimes this can be accomplished within a single publication. For example, an evaluation report might contain an executive summary, the main text with explanatory side-notes and brief tutorials, a glossary, and a technical appendix. The executive summary would be targeted to policymakers and administrators, while the main text would be written for a mixed audience. The side-notes, tutorials, and glossary would serve to familiarize non-technical audiences with basic statistical terms. And the technical appendix would provide details of the methods and results that would be most useful to other researchers. The structure of such a report allows for a great deal of information to be presented in various meaningful ways and in sections of varying complexity and detail. This helps to avoid the possible alienation of some audiences as a result of forcing non-technical readers to wade through complicated details in an effort to extract something useful. When attempting to present results in a meaningful way, researchers should consider the intended audiences and use multiplicity in reporting so that different reports and subsections of reports contain appropriate information. Other things to consider when reporting results include the intended purpose of the research, the population targeted by the study, and the expected use of the results by program developers and policymakers. Recognizing these elements is essential to effective reporting of research findings. Ideally, readers from different audiences are able to progress through the layers of results to the extent that they desire to get more specific information while remaining comfortable with the technical detail of the information presented.

An alternative to the structured evaluation report that may be even more effective for reaching multiple audiences is to produce multiple publications. These may take the form of peer-reviewed articles in academic journals, policy briefs distributed to targeted mailing lists, and editorials in newspapers and periodicals. Some may claim that this practice is unprofessional because it involves publishing the same thing in different locations. However, if secondary publications like policy briefs and editorials emphasize the citation of the original peer-reviewed journal article, then a successful attempt to reach multiple audiences may serve to disseminate important information to a large number of people, draw attention to the original academic publication, and possibly increase the visibility of evaluation research in general. What's so bad about that? I argue that this approach is essential to improving the utility of research findings for making informed decisions about policies and programs in areas where many of the stakeholders are unfamiliar and potentially uninterested in the statistical methods behind the research.

MAKING STATISTICS MEANINGFUL: SOME EXAMPLES

The sections that follow illustrate some examples for making statistics more meaningful to various audiences. The examples focus on statistics and methods commonly used in evaluation and policy research. Some of these examples are straightforward and produce interpretations that are likely to be understood by most audiences. Other examples are more complex, especially those involving errors of estimation, and may be better suited for more sophisticated audiences. When describing statistical results, it is important to recognize the information that each statistic represents, and to evaluate the importance of that information for answering the research questions and for communicating with the intended audience. Differences between demographic or program groups are often of key interest and are usually represented with regression coefficients or differences in group means. Other statistics including correlations, R^2 , and standard errors convey information about the precision of predictions or the fit of

statistical models. While it is relatively straightforward to describe simple group differences to non-technical audiences, the explanation of precision of predictions to these audiences is often neglected. Nevertheless, information about variability in expectations may be just as important to convey as differences in group means.

Descriptive Statistics in Evaluation and Policy Research

The use of statistics can generally be encompassed in two areas: descriptive statistics and relational statistics.¹ Descriptive statistics are often used to identify and gauge the severity of problems—"Less than fifty percent of students tested were proficient in reading." The purpose may be to bring a potential problem to the attention of policy makers or to outline current status on an outcome relative to a programmatic goal. These statistics help to describe the nature and extent of problems by quantifying a situation or condition. Although they sometimes imply relationships, they never quantify directly the strength of any relationship.

Descriptive statistics usually describe some aspect of the statistical distributions of policy relevant variables. Percents, proportions, averages, ratios, and many other statistics are commonly used to describe or gauge problems. These metrics are familiar to many audiences due to their common use in political surveys, weather forecasts, and other areas. These simple statistics are popular in policy and evaluation research for relating basic information and making elementary comparisons. Other descriptive statistics such as variance or skewness, which convey very important information, are not widely understood. For purposes of illustration, consider two simulated distributions of SAT scores for two different random samples of college-bound students. One sample is from very affluent high schools, and has a mean SAT score of 1100, a standard deviation of 75, and a skewness of 2.4. The other is from high schools in impoverished areas, and has a mean SAT score of 750, a standard deviation of 100, and a skewness of -1.9 .

If this information is reported in standard statistical vernacular as above, without worthwhile explanation or qualification, many potential consumers of policy research might miss the message because they don't have extensive knowledge of the properties of the statistics reported. Referring back to the guidelines for meaningfulness, the understandability and interpretability of standard deviation, variance, and skewness statistics are weak. One method for improving how information from these distributional spread and shape statistics is described relies upon an audience's knowledge of simpler statistics such as percentages. Variance and skewness measures might be explained more effectively if they are phrased in the context of distributional density, or the proportion of observations at, within, or beyond a particular range. Standard deviations do this for spread (empirical rule: 68% of observations in a normal distribution are within 1 standard deviation of the mean), but the percentile cutoff points at each standard deviation are difficult to explain. An alternative is to report an inter-quartile range: "The middle half of the SAT scores of students from poor schools lie between 760 and 860." Another alternative might be to report a restricted range—"Ninety percent of these scores lie between 640 and 900." The differences between these alternatives for representing distributional spread are subtle, and either of them would be adequate in most scenarios. The key is to state the information without using statistical "buzz words."

The distributional density method is even more powerful when relating information about skewness. The statistical parameter for skewness has no units. Its formula removes the effect of scale so that values of skew can be compared across distributions. This has the unfortunate side effect of making the skewness statistic meaningless to non-statisticians. However, the

presentation can be easily improved by recognizing the knowledge that the audience already has. For example, an alternative method for presenting information about the shape of skewed distributions is to report the distance between percentile cutoffs in the original scale of the distribution. In effect, the lengths of the tails of the distributions are expressed in the original units. For example, “In impoverished schools, the lowest performing 5% of students scored at least 180 points below their schools’ average. However, in affluent schools, this distance between the low performers and the school average is only 80 points.” This comparison helps to quantify how much longer the left tail from the impoverished schools’ score distribution is compared to the left tail of the affluent schools.

Expanding on the previous example, instead of comparing the lowest scores from two distributions, one might compare the right and left tails of the same distribution by showing the distance between the 5th and 95th percentiles and the median or mean score—“Students from affluent schools who performed worse than 95% of their peers scored at least 80 points below average, while students from these schools who performed better than 95% of their peers scored at least 140 points above average.” This is simply a comparison of the distance between the 5th and 95th percentiles and the mean. Another approach compares the value at a particular percentile to a maximum or minimum value. For example, “The scores for the top 5% of students from affluent schools had a range of 560 points (i.e., from 1,240 to 1,800), while the scores for the bottom 5% of students from these schools had a range of 270 points (i.e., from 750 to 1,020).” This is simply a comparison of the lengths of the left and right tails from the affluent schools’ distribution.

These comparisons could be represented using a figure, yet the figure would fail to relate the absolute size of the tails of the distributions in real numbers. By quantifying these differences in addition to the figure, consumers of policy and evaluation research are able to carry away information that can be disseminated and understood even when that figure is unavailable. That is not to say that these interpretations can be easily understood by anyone, but they are surely more meaningful than the traditional statistics of variance, standard deviation, or skewness.

Distributional density comparisons might be most understandable and interpretable when comparisons are made between cutpoints and the mean. However, the mean is not robust to outlying data points. Therefore, one might consider using the median in these comparisons when distributions are heavily skewed or contain influential outliers that might pull the mean away from the center of the distribution. Similarly, it is important to be consistent when choosing cutoff points so as to avoid exaggerating some information and understating other information. Obviously, the choice of cutoff points changes the emphasis of the statement. It is important to remember that the ability to make a meaningful statement is worthless if the information conveyed in that statement is misleading or inaccurate.

Relational Statistics in Policy Research

While descriptive statistics are commonly used to describe and gauge the severity of problems, relational statistics are used to describe and gauge the strength of relationship between two or more variables, or to estimate the effect of one variable on another (see note 1). Typical research questions that relational statistics can address are “How much of an impact did policy or program *X* have on outcome *Y*?” or “How strong is the association between variable *X* and outcome *Y*?”

In most policy and evaluation research, regardless of the underlying design, the primary statistic that addresses the research question can be called an “effect size” estimate. Some

purists would argue that “effect size” should be reserved for “standardized” treatment effects often reported with experimental or quasi-experimental research; however, in practice, this term is often used to describe both estimates of causal treatment effects and estimates of strength of relationships. For the purposes of this article, “effect size” is used to represent any statistic used to quantify the effect or relationship of interest. It is also important to note that this article does not address the issues of Type I and Type II errors in inferential tests (see Posavac, 1998; Rosenthal & Rubin, 1994, for meaningful interpretations of these concepts).

By far, the most widely used class of methods that produce relational statistics is linear modeling. Examples of popular linear models are analysis of variance (ANOVA) and its varieties, correlation, regression analysis, path models, and hierarchical linear models (HLM). Even popular non-linear analyses such as logistic and poisson regression are variants of the generalized linear model (Nelder & Wedderburn, 1972). As such, the concentration here will be on examples that pertain to such models, although the general themes may carry over to other analyses.

Most analytic approaches will produce a variety of statistics which can be used to represent effect size. Correlations or coefficients of determination are often more popular in non-experimental studies because they quantify the strength of relationships that exist among two or more variables of interest. Treatment parameters from ANOVA or slope coefficients from regression analysis are often preferred in experimental or quasi-experimental studies because of their straightforward representation of size and direction of treatment effects. While any of these estimates could be used in policy or evaluation research to convey the size of an effect, they may not be sufficiently meaningful to non-technical audiences. Furthermore, it may be easier to gauge the magnitude of particular effect size estimates for a particular context when they are presented in more meaningful ways. For example, effect sizes are often evaluated relative to the costs of a program, political and social values, and the availability of alternative treatments. Reporting statistics that are understandable, interpretable, and comparable can make this process a bit easier.

Making correlations more meaningful. The Pearson correlation is often thought of as the simplest statistical indicator of strength of relationship between two variables. Ironically, it might be the least interpretable statistic produced by linear models. This problem is often made worse because most researchers simply report the value of the statistic without much explanation. For potential consumers of policy research, knowing the actual size of the correlation is worthless unless they are familiar with correlations and their properties. It does help to explain the range of possible values for correlations (i.e., -1 to 1), and to explain other characteristics of correlations. However, even the research community acknowledges that there is considerable subjectivity in determining what large, moderate, or small correlations look like, and that the practical magnitude of a correlation can depend heavily on context (Glass, McGaw, & Smith, 1981; Rosenthal & Rubin, 1982).

The problem of interpretability of the Pearson correlation is so pervasive, that even a large proportion of the research community is unaware of its mathematical interpretations. While most researchers are aware that a correlation is the positive or negative square root of the proportion of common variance between two variables, it also shows the expected standard deviation change in one variable per standard deviation change in the other (i.e., the traditional standardized effect size for two continuous variables). A third interpretation of the Pearson correlation is that it equals the cosine of the angle of separation between regression lines of Y on X and X on Y . Extending the bivariate correlation, partial correlations also control

for confounding variables, providing the same type of estimate after holding the partialled variables constant. Still, these interpretations are not very helpful to the non-statistician.

The presentation of scatter plots can help to facilitate the conveyance of information represented by correlations; however, this produces a dependence on the figure, and the ability to communicate exists only when the figure is available. Rosenthal and Rubin (1982) developed an alternative means for relating effect size from a correlation that is understandable, interpretable, and comparable. They termed this new method the binomial effect size display (BESD). It involves dichotomizing both variables around cutoff points that divide values considered to be successes from those considered to be failures. More specifically, the BESD is calculated by dichotomizing both variables around their medians (i.e., values below the 50th percentile are recoded to 0, while values above the 50th percentile are recoded to 1). Then the dichotomized variables are tabulated in a two-way contingency table. Therefore, the BESD shows the proportion of observations above the 50th percentile on one variable given that they are above the 50th percentile on the other variable. Figure 1a would be interpreted by stating, "Individuals above the median on *X* had a 66% success rate, while those below the median on *X* had a 34% success rate."

Unfortunately, use of this method does impose assumptions on the data that are likely to be inappropriate. Thompson and Schumacker (1997) argue against the validity of the BESD, and although it does add interpretability to a correlation, it can distort results. The key problem is that the cutoffs under the proposed BESD are always set equal to the variables' medians. If these cutoffs are inappropriate, the effect size estimate will be wrong. This is because the BESD assumes that the probability of "success" underlying each variable is .5. This may be an arbitrary cutoff in practice. However, the general approach of the BESD is promising due to its transformation of a correlation into a simple 2×2 contingency table.

An improved version of the BESD involves specifying cutoffs for each variable that are based upon prior knowledge or theory. It is important that the researcher's choice of cutoffs be easily defensible; any discretion on the part of the researcher will have to be supported. The best cutoffs are those that are pre-established. Examples of such cutoffs include the fail score on an exam, the number of unexcused absences allowed by a school, income at the poverty level, and the legal limit of a driver's blood alcohol level. For illustrative purposes, consider a correlation of $-.40$ between scores on a graduation exam and the number of absences for a sample of 3,000 students. The failure rate on the test is 10%, and almost 40% of all students have exceeded the limit for unexcused absences. As shown in Figure 1b, the new version of the BESD converts the continuous data into a 2×2 contingency table where the test failure rates can be compared to number of absences. Figure 1b could be interpreted by stating, "The failure rate on the test is 5.6% for students who did not exceed the absences limit and over 17% for those who did."

Although the primary analyses of the relationship between absences and test scores in this example involved calculation of a traditional correlation, the ability to dichotomize both variables in a BESD-style produces a more meaningful interpretation of the results than the correlation alone. This alternative method of reporting can be understood without any knowledge of correlations, and its dissemination is not dependent on the presence of a visual aid.

Making r^2/R^2 more meaningful. Coefficients of determination are another set of indicators of strength of relationship and are usually reported with regression analyses and sometimes with ANOVAs as an indication of strength of relationship. The statistic has various forms including the bivariate r^2 , which does not control for potential confounders; the multiple R^2 ,

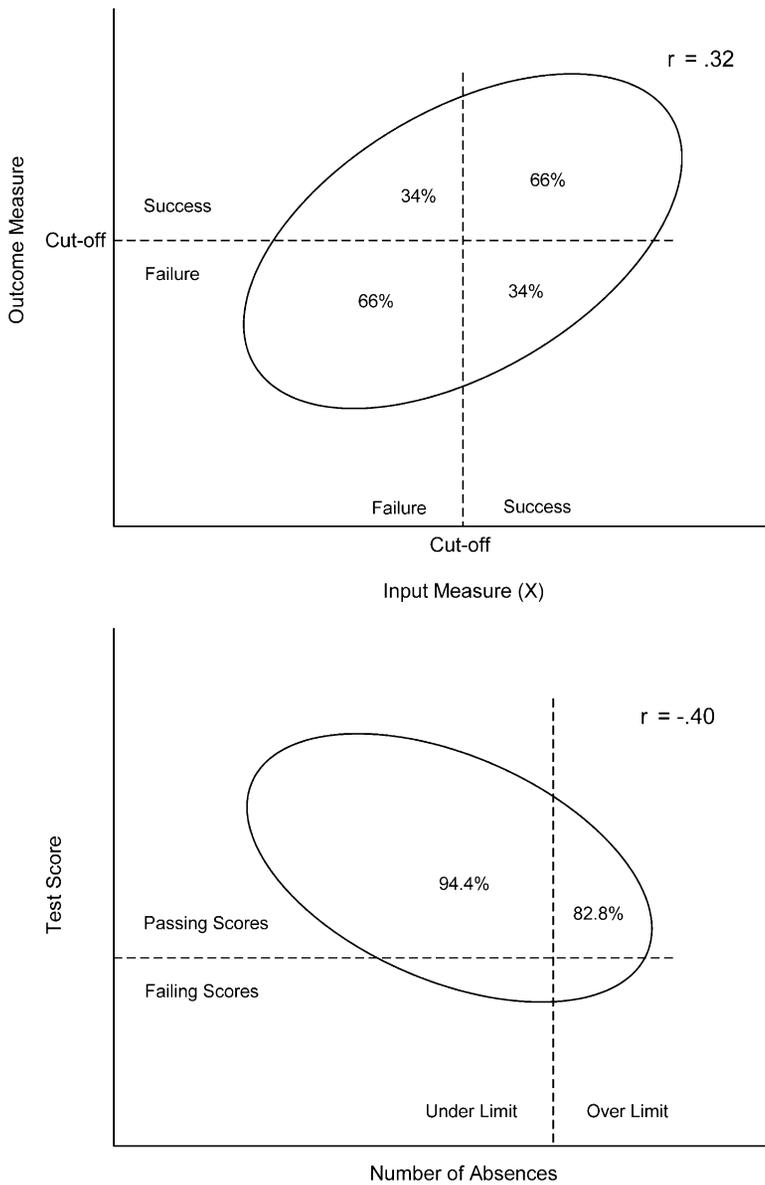


Figure 1. Illustration of traditional (a) and improved (b) binomial effect size displays.

which conveys strength of relationship between a set of predictors and an outcome; and the partial or incremental R^2 , which shows the added contribution of one or more predictors after controlling for confounders. Most often, they are interpreted as representing the proportion of variance in an outcome that can be explained by one or more predictor variables. Obviously, this interpretation assumes familiarity with the concept of variance; and unfortunately, like the Pearson correlation, there is a great degree of subjectivity in deciding whether this statistic is

small, moderate, or large. More importantly, its interpretation still involves the use of statistical jargon, and the effect size is not expressed in an interpretable metric.

The key to getting beyond this problem and truly making the coefficient of determination more understandable and interpretable is to recognize the kind of information it provides; that is, the accuracy of the predictions of the model, or how tightly the observed data hugs the regression line or space. Unfortunately, conveying this information to non-technical audiences can be quite a challenge. The examples here are more interpretable than r^2 or R^2 ; however, they still require that the audience have some knowledge of statistics. It may be the case that these methods are most effective at conveying information about model fit to researchers and technical audiences, while avoiding the pitfalls of the “variance explained” concept. Arguably, “variance explained,” “measurement error,” and “prediction error” are not as easily understood as other estimates of effect size.

When the audience includes non-technical readers, the information from bivariate and multivariate coefficients of determination can be re-expressed in terms of the spread of the observed data points around the regression line (i.e., the error variance) using the raw metric of the outcome variable. For example, to illustrate the strength of relationship between family wealth and academic achievement, a researcher might state, “Approximately half of the students at the 25th percentile of wealth have achievement scores between the 15th and 55th percentile while about half of the students at the 75th percentile of wealth have achievement scores between the 41st and 82nd percentile.” This interpretation provides a meaningful quantification of the overlap in achievement scores for students at different poverty levels. This overlap is not nearly as clear when the only information presented is an r^2 of .30 or a correlation of .55.

It is important to recognize that the interpretation above, like any distributional density approach, is heavily influenced by the cutoffs chosen. This is especially important when describing relationships between two variables, where it is essential that equally sized intervals be used for both variables. In the example above, the interquartile range was used for both variables (i.e., the values at the 25th and 75th percentiles of the achievement distribution at the 25th and 75th percentiles of wealth). If a wider range was used for wealth than for achievement, the overlap in scores would be understated. If a wider range was used for achievement than for wealth, the overlap in scores would be exaggerated.

Alternatively, if the intended audience was generally familiar with regression methods, the researcher could simply report the square root of the mean squared error (RMSE), which is sometimes interpreted as an indicator of the amount of “noise” around the model. Because RMSE is always in the raw units of the dependent variable, it can be directly compared to the standard deviation of the dependent variable (Y). If the assumption of constant variance is correct (this assumption accompanies all linear models), then RMSE is a consistent estimator of the standard deviation of the observed data around the regression line. In other words, RMSE shows the standard deviation of the conditional distribution of Y given X , where Y is the dependent variable and X is an independent variable. Therefore, a 95% confidence band for the observed values of Y at any one value of X would be approximately $2 \times$ RMSE wide.

An alternative approach that conveys information about the fit of the model capitalizes on the properties of statistical distributions when interpreting RMSE. This method is appropriate for audiences that understand regression and might be helpful for less technical audiences if sufficient explanation is given. This approach also helps to provide a clearer picture than R^2 of just how accurately the model predicts the data, given that RMSE is expressed in the raw units of the dependent variable. A proven fact in statistics states that when any distribution

(X) is multiplied by a constant (C), then the standard deviation of the resultant distribution (Y) is C times as large as the standard deviation of X . However, the variance of Y is C^2 times as large as the variance of X , even though the spread distribution has really only increased by a factor of C , not C^2 . This is evident in the fact that the range of Y is exactly C times as large as the range of X . This is very important for the purposes of interpretation. As such, for a linear model, with an $r^2 = .80$, where the standard deviation of the dependent variable is 100 points, and the RMSE is 45 points, one may calculate $1 - (\text{RMSE}/\text{SD}(Y)) = 1 - (45/100) = .55$, which can be interpreted as, "The expected range of the dependent variable decreases by 55% after controlling for X ." In the previous example of wealth and achievement, with an $r^2 = .30$, the standard deviation of achievement is 100 points, and the RMSE is 84 points, showing that, "The expected range of achievement decreases by 16% after controlling for wealth."

Note that RMSE can also be used to calculate the proportion of variance that is not error, $(100^2 - 84^2)/100^2 = .2944$, which is very close to the r^2 value of .30. However, the r^2 statistic does not represent the strength of the relationship nearly as well as RMSE for two reasons. First, unlike RMSE, r^2 has no units, so the layperson is likely to accept someone else's opinion of small, medium, or large. Second, r^2 is calculated on a variance scale (i.e., as the proportion of variance explained), so it can only be gauged relative to the distribution of the dependent variable in squared units (i.e., the units of variance). This is exactly why r^2 may appear large to some researchers, even when the model isn't a tight fit to the data. Reporting and explaining RMSE in addition to, or even instead of, r^2 or R^2 will help to avoid overstating what may be truly small effects.

Another approach to improving R^2 incorporates information about the assumed distribution of the errors in the model. For example, a method for linear models involves the use of the properties of the error distribution, which is assumed to be normal. Simply put, the researcher could report the observed proportion of students with given characteristics scoring above or below average. For a given distribution of scores normally distributed around the regression line at a chosen value of X , a researcher could estimate the proportion of scores above the overall mean of Y . To compute this proportion for a given value of X (when the sample is large), one calculates the predicted value of Y at X , then subtracts the mean of Y and divides by the RMSE. This produces a z -score that can be compared to the normal table in table in the back of most statistics books. From the previous example, "For students near the 75th percentile of wealth, approximately 63% have above average achievement scores, while only 33% of students near the 25th percentile of wealth have above average achievement scores." This clearly illustrates the effect of wealth while also noting the degree of overlap in the distribution of scores. If the r^2 for the model were higher, the two proportions would be further apart, suggesting less overlap in these two distributions and a tighter fit of the data around the regression line. A visual representation of this approach is shown in [Figure 2](#).

This method is somewhat similar to the BESD approach for correlations. However, with this technique, the cutoff for Y must be equal to the predicted value at the average of the two values of X in order to avoid misrepresenting the overlap in the distributions. Using the 25th and 75th percentiles for X and the mean of Y as cutoffs is generally prudent. This method is also similar to [Smith and Glass' \(1977\)](#) approach to converting standardized effect sizes to percentages; however, their method produces only one estimate showing the proportion of individuals that had outcomes greater than the mean of the control group. As [Myers \(2001\)](#) points out, this one-sided interpretation can inflate the apparent size of treatment effects. The method proposed here avoids this by making comparative statements at more than one level of the predictor variable (e.g., at the 25th and 75th percentile of the independent variable).

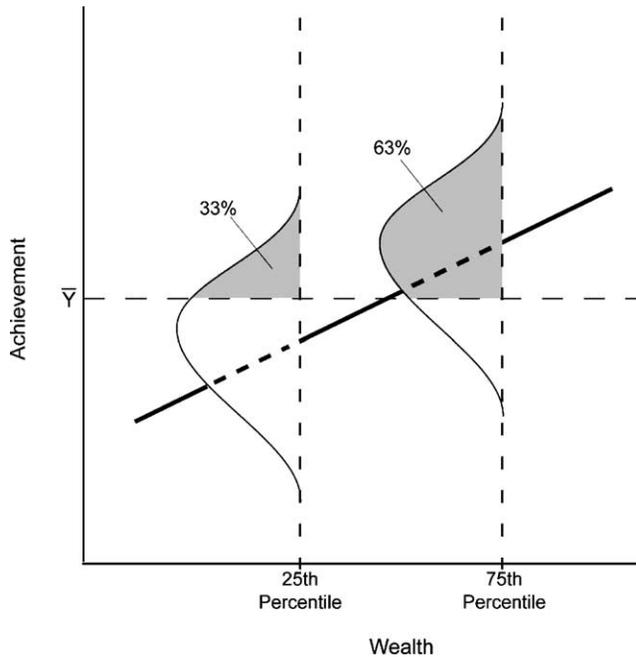


Figure 2. A Visual representation of an alternative method for representing effect size and strength of relationship between wealth and academic achievement. The normal curves represent the distribution of data points along the dotted lines at the 25th and 75th percentiles. The shaded areas represent the proportion of scores above the overall mean, \bar{Y} , when holding wealth constant at the 25th and 75th percentiles.

Making regression coefficients more meaningful. While correlations and coefficients of determination are used to indicate strength of relationships, regression coefficients and group mean differences from ANOVA are better suited as estimates of treatment effects. These statistics usually have excellent interpretability and understandability. They represent the expected change in an outcome per unit change in a predictor. For example, a regression coefficient of .7 might be interpreted as, “An increase of 10 points on the motivation scale is associated with a 7 point increase in test score.” If multiple regression or ANCOVA is used, then the estimates produced are adjusted to control for confounding variables included in the analysis.

If the predictor variable is a dichotomy, as is the case in many controlled program evaluations, then the raw regression coefficient is an estimate of the difference in outcomes for those who participated in the program and those who didn't. If the method of analysis is ANCOVA, then the treatment effect is the difference between adjusted group means. In either case, describing treatment effects for categorical variables is relatively understandable when they are expressed in the raw metric of the dependent variable. For example, “After adjusting for differences associated with race and socioeconomic status, the estimated impact of the program is a 20 point increase in reading scores.”

However, many evaluation researchers often place undue emphasis on the comparability of the statistics they report, usually to the detriment of understandability and interpretability.

A common practice is to standardize the dependent variable prior to analysis and report all estimates in standard deviation units. This leads to lower understandability and interpretability of the results, thereby reducing the utility and visibility of the research to policy-makers and the public. Although standardized effect size estimates are an important product of policy and evaluation research, they should always be accompanied by more understandable and interpretable statistics. In most cases, reporting the raw treatment effect estimate along with a standardized effect estimate is sufficient.

While the above argument pertains only to comparisons of discrete groups, effective representation of effect sizes for continuous independent variables can be much more complicated. The key problem with a raw regression coefficient for a continuous variable is that its magnitude depends heavily on the scale of its predictor variable. A common example involves the use of salary as a predictor variable. If salary is included in a regression analysis, and its scale is in dollars, then the raw regression coefficient represents the expected change in the outcome per one-dollar increase in wealth. This is likely to be an infinitesimal amount, making the raw coefficient uninformative. Common practice is to scale variables such as salary in larger numbers of units (hundreds, thousands, etc.). This improves the meaningfulness of the raw coefficient, and again we have good understandability and interpretability, but the raw coefficient still cannot be compared to the coefficients for other variables that are not on the same scale.

Unfortunately, producing regression coefficients that are comparable, while maintaining good understandability and interpretability is not so straightforward. The traditional approach is to calculate standardized regression coefficients, which represent the expected number of standard deviations change in an outcome per one standard deviation change in a predictor. If the regression model only has one predictor, then the standardized regression coefficient is equal to the Pearson correlation between the predictor and outcome variables. While this method produces statistics that are comparable, even across different studies and outcome variables, the fact that they are scaled in standard deviation units reduces understandability and interpretability.

This can be avoided to some degree by using "partially standardized" coefficients which represent the change in the outcome variable (in raw units) per standard deviation change in the predictor. They are calculated by simply multiplying a raw regression coefficient by the standard deviation of its predictor variable. This would yield an estimate of the change in the outcome, in its original metric, associated with moving one standard deviation on the predictor variable. For example, "A one standard deviation increase in socioeconomic status is associated with a 35 point increase in reading score." These estimates are comparable for any model with the same dependent variable, and they retain more understandability than fully standardized estimates. Alternatively, avoiding the standard deviation altogether, a researcher could multiply a raw regression coefficient by the corresponding interquartile range of the predictor variable. This would yield an estimate of the change in the outcome associated with moving from the 25th to the 75th percentile on the predictor variable. For example, "Students at the 75th percentile of wealth scored an average of 150 points above students at the 25th percentile of wealth."

In a non-linear analysis like logistic regression, similar techniques can be used to rescale regression coefficients. The only difference here is that the outcome is the log-odds of an affirmative response on the dependent variable, and regression parameters must be exponentiated in order to have direct meaning as odds ratios. Like coefficients from the traditional linear model, the odds ratios reported with a logistic regression often need to be compared.

Raw parameter estimates can be multiplied or divided in much the same way as normal-model coefficients, then transformed through the formula $100(e^{\beta} - 1)$ to produce final estimates (i.e., percent change in odds) that are comparable. For example, “A one standard deviation increase in motivation is associated with a 62% increase in the odds of scoring at or above proficient.”

Another approach which is appropriate for both linear and non-linear models is to report predicted values for observations with specific values of the dependent variables. For example, predicted outcomes for prototypical groups or individuals can be estimated by substituting typical or average values of the independent variables into a regression equation. If the analysis involves a continuous outcome, the predicted values will be expressed in the raw units of the dependent variable, “The predicted score for an average urban school is 430, while the predicted score for an average suburban school is 510.” Likewise, if the analysis involves a categorical variable, the predicted values should be expressed in probabilities or proportions. These are often very helpful for linear and non-linear regression models with significant interactions between predictors. For example, a model where the use of reading coaches is especially effective in urban schools might show that, “On average, urban schools are predicted to have a 55% proficiency rate in reading. However, urban schools with reading coaches are predicted to have a 75% proficiency rate.”

For every analysis involving linear or non-linear models, researchers must attend to all three guidelines for meaningful reporting of statistical results. Given that the utility and visibility of policy and evaluation research is directly influenced by the breadth of the audience it reaches, understandability and interpretability should never be compromised as a result of efforts to gain comparability. If comparability is a primary goal, and standardized effect sizes are used, then raw effect estimates should be presented as well to preserve understandability and interpretability.

CONCLUSIONS

Methods for improving the meaningfulness of effect size statistics range from the very simple to the very complicated. A researcher might have to implement several computational steps in order to reformulate a statistic adequately. However, the benefits surely justify the effort. Regardless of the number of steps required to present information in more familiar terms, if the researcher takes the time to perform these steps, then the consumer of policy research has more time to ponder the implication of the results instead of trying to figure out how to interpret the statistical results. Ultimately, if policy and evaluation researchers strive to report results in ways that expand the potential audience and increase the visibility of research, the ability to inform the public is improved, and the likelihood of implementing promising programs and policies is increased.

The examples presented here by no means exhaust the possible alternatives for presenting statistical information in policy research. However, they do illustrate some general approaches to making statistics meaningful. The first of these is to consider the basic nature of the information that needs to be conveyed. What is the most fundamental statement that can be made which adequately relates the quantitative results? The answer to this question should state more than just “higher” or “lower.” It should quantify the effects or relationships in a way that is understandable and interpretable to multiple audiences, while also enabling direct comparisons to other effects or relationships. This requires intuition and creativity. Ideally, the researcher would reformulate or develop a statistic that discloses the magnitudes of effects or relationships in raw units. This may simply be the raw metric of the dependent variable.

However, the chosen unit should be familiar to the vast majority of people who might use the information. Once this is achieved, then the researcher can consider rescaling the statistic so that effect sizes can be compared across factors within a study and possibly with effects from other studies on other outcomes.

The suggestions outlined here should not be taken as arguing the abandonment of traditional statistics. Alternative methods for presenting statistical information should be considered as supplements to traditional statistics in the context of policy and evaluation research. Despite the esoteric nature of many of our traditional statistics, they have desirable statistical properties that most of these alternative approaches do not. Therefore, researchers should retain traditional statistics in their tool kits, but augment them with creative means for disseminating important information to different audiences.

NOTE

1. Some may divide statistics into descriptive and inferential statistics. This distinguishes statistics based upon purpose: descriptive statistics provide point estimates, while inferential statistics serve to test hypotheses. Here, I choose to differentiate statistics based upon the information they convey: descriptive statistics quantify a situation based upon a single measure, while relational statistics quantify relationships between multiple measures or conditions.

REFERENCES

- Cohen, J. (1988). *Statistical power analysis for the social sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Hemphill, J. F. (2003). Interpreting the magnitudes of correlations. *American Psychologist*, 58, 78–79.
- Myers, D. G. (2001). *Psychology*. New York: Worth Publishers.
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, 135, 370–384.
- Posavac, E. J. (1998). Toward more informative uses of statistics: Alternatives for program evaluators. *Evaluation and Program Planning*, 21, 243–254.
- Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166–169.
- Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science*, 5, 329–334.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752–760.
- Thompson, K. N., & Schumacker, R. E. (1997). An evaluation of Rosenthal and Rubin's binomial effect size display. *Journal of Educational and Behavioral Statistics*, 22, 109–117.