Evaluation FastFacts

from the Evaluation Center@HSRI

Volume 2, Issue 4 June 2003

This is one in a series of briefings on new and current mental health services evaluations, resources, and methods. We hope FastFacts will be a quick and easy way for you to learn important information in the field of evaluation. If you have any ideas on how FastFacts could be more useful to you, please contact Dow Wieman, Ph.D. at 617-876-0426 x2503 or dwieman@hsri.org.

Not All Statistically Significant Differences Are Alike: Equivalence Analysis in Mental Health Services Evaluation

Researchers and evaluators in the field of mental health care and psychosocial services typically use statistical tests to estimate whether different programs or systems of care are likely to produce different results. These may be differences in satisfaction, symptom relief, quality of life, or some other outcome for individuals in the programs or systems. But policy makers, providers, administrators and others who use the results of research and evaluation are often less concerned about whether the results of the two programs are different than about whether or not results are similar. This is a different question.

The Director of Planning in a state mental health agency may want to know, "Will a particular innovative program that is less expensive produce results similar to those of a program already in place?" The director of an agency funding a project comparing two psychosocial interventions may ask, "Do different racial or ethnic groups have equivalent outcomes in each?" A program developer evaluating a new intervention may ask,

continued on page 2

Inside This Issue

The Evaluation Center@HSRI

is a technical assistance center funded by the federal Center for Mental Health Services (CMHS), Substance Abuse and Mental Health Services Administration (SAMHSA), and operated by the Human Services Research Institute (HRSI). The mission of the Center is to provide evaluation technical assistance to state and non-profit and private entities including, but not limited to, consumers, families and provider groups. The Center presently has six programs designed to fulfill this mission—

- Conferences & Training
- Consultation Program
- Knowledge Assessment & Application
- Multicultural Issues in Evaluation
- Toolkit & Evaluation Materials
- Topical Evaluation Networks & Web

For more information on the Center, please visit our website at:

http://www.tecathsri.org

Or contact us at:

Tel: 617.876.0426

Fax: 617.497.1762

Email: contacttec@hsri.org

Address:

2269 Massachusetts Avenue
Cambridge, MA 02140

— continued from page 1

"If the new intervention achieves outcomes that are better than what are currently obtained, will they be enough better to consider them significant for clinical or policy purposes?" A policy maker may want to know, "If a managed care program were implemented, would outcomes for persons with serious mental illness be at least equal to those for a comparable group under fee for service?"

Technically, in conventional statistical tests of difference, the investigator tries to reject the null hypothesis of no difference between the mean scores on some measurement of the two groups, and when failing to do so, asserts that a difference between the groups cannot be proven. The individuals in the examples above, however, wish to do what conventional difference testing says is impossible: in effect, they want to prove the null hypothesis of no difference. Is there any way the evaluator or researcher can answer the question most important to these people: "Are the two things alike in their effect, and if so, how much alike?"

In fact, there is such a method. Known as equivalence analysis, it is widely used in other fields, research and evaluation (Rogers, Howard et al. 1993; Stegner, Bostrom et al. 1996; Hargreaves, Shumway et al. 1998).

Equivalence testing is based on the method of bioequivalence testing, used by the FDA and the pharmaceutical industry for determining whether a new drug is acceptable as an alternative to one previously approved. In this context, equivalence analysis is a method for estimating whether a difference between two groups, if one exists, is small enough (according to some pre-determined threshold) to warrant considering the results as equivalent for clinical or policy purposes.

Equivalence testing differs from traditional hypothesis (difference) testing in that it reverses specification of the null and alternative hypotheses (Hargreaves, Shumway et al. 1998). In difference testing, the null hypothesis is that differences among group means are zero. The alternative hypothesis is that these differences are not zero. In equivalence testing, the null hypothesis is that the difference among group means is greater than some minimal difference representing practical equivalence. The alternative hypothesis is that the difference is not greater than this specified minimum difference.

In the analysis of differences among groups such as those described in the above examples, this step allows the researcher to estimate whether identified significant differences are meaningful differences for policy makers or clinicians. Equivalence analysis also makes it possible to determine whether non-statistically significant differences may be the consequence of small sample sizes and/or large variability rather than actual equivalence between the two programs or systems.

Equivalence analysis, effect size, power analysis, and null hypothesis tests are research tools, each having its own purposes and limitations. Null hypothesis testing is the grandfather of all statistical tests. It is a well established method of determining if a significant difference exists between two populations, but the test only shows the existence of a difference—it doesn't describe it. Effect size fills in the gap for null hypothesis testing by describing the difference between populations, but an effect size alone does not tell if a difference is significant. Power analysis is used to determine the ideal number of cases to give diagnostic results to a study; without sufficient power statisticians would not be able to find significant differences or similarities (and with "excessive" power, the differences, though statistically significant, will be trivial for practical purposes). Equivalence analysis can be thought of as the complementary test to null hypothesis test-

continued on page 3

— continued from page 2

ing; it shows the existence of statistical equivalence between two populations, but requires effect size to describe the variation between the groups.

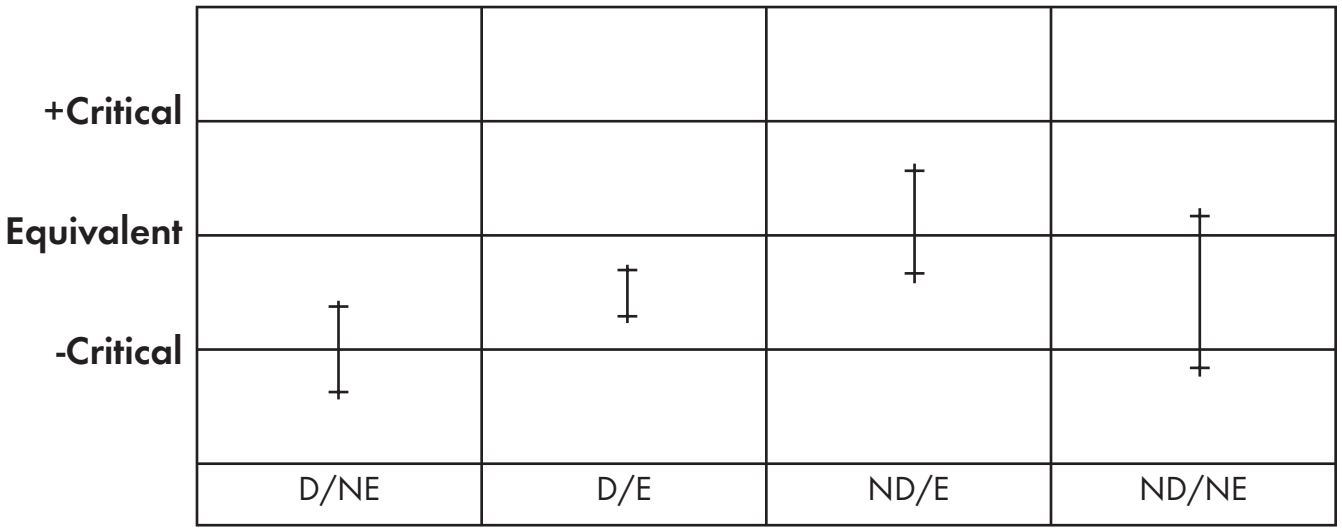
The researcher establishes equivalence boundaries for the effect size and then determines equivalence or non-equivalence by calculating a confidence interval for the test case. The conventional standard for accepting bioequivalence of drugs (analogous to the 95% confidence interval in difference testing) is that the group mean for the test drug on some outcome, for example, blood plasma uptake, is within a small enough range of the group mean for the previously established drug (the control group) that the difference is not considered substantively important. The range used for the equivalence comparison varies by case. The best accepted method is to use a percentage of the mean of the control group; however, a percentage of the standard error of the control group or of difference between the two groups may also be used, or an ad-hoc range determined by previous experimentation.

Because no standard equivalence range has been established for policy relevance of differences between managed care and fee for service (or other psychosocial treatment interventions) comparable to the range for drug trials, the analyst may wish to take a more descriptive approach, computing multiple ranges of 5, 10, and 20 percent of the mean of the control group. Thus, a 5 percent difference between group means would be the most conservative standard for establishing equivalence, and 20 percent would be the most liberal.

Difference testing and equivalence analysis are not mutually exclusive. Performed together, they yield four possibilities:

1) Different and non-equivalent (D/NE)—there is a difference, and it is sufficient to have clinical or policy relevance);
2) Different but equivalent (D/E)—there is a difference but it is trivial, i.e.; the study is overpowered;
3) Not different and equivalent (ND/E)—the two conditions are indistinguishable;
4) Not different but also not equivalent (ND/NE)—the variability is too great relative to the effect size to interpret, i.e.; the study is underpowered.

The four possibilities are illustrated below in a graph of comparisons using four convenient pairs of data. Confidence intervals of the test group are



continued on page 4

— continued from page 3

shown as vertical bars. The entire length of the vertical bar represents the 95% confidence interval used in the T-test to test if the groups are significantly different. If 0 is within this confidence interval the two groups are not statistically different (ND). If 0 is not contained then the groups are statistically different (D). The equivalence test uses a 90% confidence interval; this is shown in the graph on the same line as the 95% confidence interval but it is smaller so its boundaries are denoted by hash marks. If the entirety of the 90% confidence interval is bound within the critical values for equivalence then the two groups are statistically equivalent (E). If any part of the 90% confidence interval lies on or outside of that boundary then the two groups are not statistically equivalent (NE).

Note that the power greatly affects the outcome of the analysis. For D/E the excess power effectively shrinks the test group's confidence interval, while in ND/NE the lack of power effectively grows the test group's confidence interval. Investigators using equivalence analysis for comparing programs or systems will need to address two issues. First, much greater power is required to determine equivalency than to determine difference. Consequently, it is likely that most existing studies of mental health services are underpowered for this purpose. Second, the mental health field will need to devise methods, comparable to the FDA's, for determining equivalence boundaries.

With these non-trivial ssues resolved, and using a format like the graph above, the researcher or evaluator may find that equivalence analysis allows them to compare groups and present the results in a way that audiences such as those in the examples would find easy to grasp, intuitively meaningful, and useful in practice.

More detailed and technical information about equivalence testing in psychosocial and services research may be found in the following references. The reader is also referred to the Evaluation Center@HSRI website, www.tecathsri.org. Coming soon: SPSS code for conducting equivalence analysis. For more information on these resources, please contact Dow Wieman at dwieman@hsri.org.

REFERENCES

Hargreaves, W., M. Shumway, et al. (1998). **Cost-Outcome Methods for Mental Health**. San Diego, Academic Press.

Managed Care and Adults with Serious Mental Illness Study Team (2001a). Managed Care and Vulnerable Populations Study Core Paper 1: Sample Survey Study, Substance Abuse and Mental Health Services Administration (SAMHSA).

Rogers, J., K. Howard, et al. (1993). "Using significance tests to evaluate equivalence between two experimental groups." **Psychological Bulletin** 113: 553-565.

Stegner, B., A. Bostrom, et al. (1996). "Equivalence testing for use in psychosocial and services research: an introduction with examples." **Evaluation and Program Planning** 19(3): 193-198.59 (4), 533-540.

continued on page 5

Change Your Subscription Status

To change your fax number or to cancel, check one of the following, and fax this page to: 617.497.1762

☐ Cancel my subscription to FastFacts

☐ Change my fax number

My name:

My new fax number:

Obtain FastFacts electronic version

To receive an electronic version of this FastFacts, go to the Toolkits & Materials section of the Evaluation Center's website at:
http://www.tecathsri.org/materials.asp.

Click the "order FastFacts" link near the bottom of the page and enter "**ff-11**" in the search box to obtain a free copy.

Receive TECNews via the Internet

Signup for TECNews, the electronic newsletter of the Evaluation Center@HSRI. The newsletter presents updates to TEC's website, product download information and current news in the world of mental health evaluation.

To signup for TECNews visit our website at:

http://www.tecathsri.org