DOCUMENT RESUME

ED 437 412 TM 030 564

AUTHOR	Hanes, John C.; Hail, Michael
TITLE	New Indicators for Program Evaluation.
PUB DATE	1999-11-06
NOTE	23p.; Paper presented at the Annual Meeting of the American Evaluation Association (14th, Orlando, FL, November 2-6, 1999).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Change; *Evaluation Methods; *Hypothesis Testing; Measurement Techniques; Program Effectiveness; *Program Evaluation; *Statistical Significance
IDENTIFIERS	*Educational Indicators

ABSTRACT

Many program evaluations involve some type of statistical testing to verify that the program has succeeded in accomplishing initially established goals. In many cases, this takes the form of null hypothesis significance testing (NHST) with t-tests, analysis of variance, or some form of the general linear model. This paper contends that, at least for program evaluation, the focus of NHST and its recommended alternatives misses the target of what evaluators really need to know about a program's success: (1) how meaningful to the client the changes attributable to the program are; (2) how many participants actually achieved these changes; and (3) how practical the program was. An approach is recommended that borrows from approaches used in the medical sciences to accumulate replicative evidence from repeated applications of the same program or from programs of a similar nature. This allows a defensible descriptive analysis of program effectiveness and efficiency. Taking a more descriptive approach means examining sustainable clinical improvement that results from the program. Using sustainable clinical improvement and the indicators of practical significance produces answers and makes statistical inference welcome, when warranted, but not necessary. (Contains 5 figures and 54 references.) (SLD)

New Indicators for Program Evaluation

John C. Hanes

Michael Hail

University of North Carolina – Greensboro

jchanes@uncg.edu

Paper presented at the annual conference of the American Evaluation Association

Orlando, FL

November 6, 1999

New Indicators for Program Evaluation

John C. Hanes and Michael Hail

Many program evaluations involve some type of statistical testing to verify that the program has, indeed, succeeded in accomplishing initially established goals. In many cases this takes the form of null hypothesis significance testing (NHST) with either t-tests, analysis of variance (ANOVA), or some form of the general linear model (GLM). Commonly utilized in the social, behavioral, and health sciences as well, such approaches have received increasing criticism from methodologists in their respective communities for many years. Whole volumes carry the debate (Harlow, Mulaik, & Steiger, 1997; Morrison & Henkel, 1970), and some journals devote single issues to the topic (The Journal of Experimental Education, 1993, Vol. 61, No. 4). Several classic articles provide reasons for extreme caution when using statistical inference – see Carver (1978), Meehl (1978), and Cohen (1994). Incidentally, Carver and Meehl lament the continuing reliance on significance testing in more recent articles (Carver, 1993; Meehl, 1997).

This paper contends that, at least for program evaluation, the focus of NHST and its recommended alternatives misses the target of what we really need to know about a program's success. Finagle's New Laws of Information state that: 1.) the information we have is not what we want, 2.) the information we want is not what we need, and 3.) the information we need is not available (Peers, 1978). Often, in program evaluation, we have what we say we want, statistical significance testing, which is usually not what we need for several reasons. Infrequently, we have confidence intervals, power analyses, effect sizes, and meta-analysis. While these move closer to what we need, they still fail to tell us three things: how meaningful to the clients were the changes which might be attributable to the program, how many program participants actually achieved these changes relative to a comparison group, and how practical was the program. Fortunately, this information is available.

The various sciences form a rough hierarchy of research rigor with regard to the selection and assignment of subjects, the control of subjects and variables, the type and range of treatments, the precision and accuracy of instrumentation, and the enforceability of protocols. Generally, the physical sciences supercede the biological, medical, and health sciences, which, in turn, have somewhat more credibility in these regards than the behavioral and social sciences. Evaluation occupies a place among the latter group.

When analyzing evidence for the effect of an experimental manipulation in the physical sciences, a mathematically predictive theory, strong instrumentation, and extensive, varied, external replication (a cumulative process) make the argument without need of statistical significance testing (Meehl, 1978). "Replicated results automatically make statistical significance unnecessary" (Carver, 1978). The medical sciences look to the randomized controlled trial, or RCT (The Standards of Reporting Trials Group, 1994), with a large sample size (Moore, Gavaghan, Tramer, Collins, & McQuay, 1998) as the gold standard for determining treatment effects, and this provides the foundation for various tests in terms of statistical significance, confidence intervals, and clinical efficacy. The behavioral and social sciences seemingly seek to emulate the medical sciences without, for the most part, the imprimatur of the RCT and sufficiently large sample sizes. When the social sciences occasionally imitate the physical sciences by utilizing replication, they often employ internal replication in the form of crossvalidation, jackknife, or bootstrap procedures. This dependence on a single sample tends to inflate estimates of replicability (Thompson, 1993).

For program evaluation, we propose borrowing several approaches from the medical sciences while dropping any unjustified inferential associations and using these approaches as a springboard for accumulating replicative evidence from repeated applications of the same program or from programs of a similar nature, much like the physical sciences do. Utilizing this combination allows a defensible descriptive analysis of program effectiveness and efficiency, both individually and in a comparative mode; our evaluation moves to a more exploratory orientation accompanied by numerical, counting, and graphical detective work (Tukey & Tukey, 1988).

Statistical Significance Testing

In program evaluation, the warrant for using statistical significance testing often breaks down on basic assumptions underlying the methodology. Many times a program lacks both random selection and random assignment. Without one or the other, comparison of test statistics with a reference distribution carries little meaning because the theoretical mathematical curve of the reference distribution is generated with the assumption of random sampling (Shaver, 1993). "...accurate significance testing *requires* randomization (random sampling or assignment) to be interpretable" (Biskin, 1998).

Behavioral science, social science, and program evaluation data often fail to meet assumptions of independence, equal variance (homoscedasticity), and normal distribution for the error term of the dependent variable(s). As Stevens (1996) points out with reference to ANOVA and MANOVA, violation of the independence assumption has serious consequences and occurs quite often. If interventions involve interactions among the individuals receiving treatment, then independence is at risk, and random sampling or random assignment does not solve the problem. Abuse of the independence assumption in the social sciences usually involves a misunderstanding about the unit of analysis, occurs frequently, and often escapes the scrutiny of a field's top journal editors (Hykle, Stevens, & Markle, 1993).

While re-expressions or transformations may help to some extent for unequal variances and non-normality, violations of these assumptions dictate caution in interpretation of the test results. Certain statistical tests exhibit robustness in the face of assumption violations, but discerning the actual intervention or treatment effects becomes increasingly difficult as the experimental design advances in complexity (Biskin, 1998).

Recognized variables sometimes suffer poor quality and reliability in their application, and a plethora of additional variables, some spurious and some of consequence, either make an unnecessary appearance in the research model or fail to receive adequately measured attention during program administration. Such measurement and specification errors can produce disastrous results for the evaluation (Pedhazur, 1997), and, as in the case of the independence assumption, even elite journals overlook the absence of important measurement components in their published studies (Whittington, 1998).

Threats to internal and external validity may directly or obliquely effect statistical significance, and program evaluation provides a rich target for such problems. Systematic error or bias may enter the evaluation via such threats, and amelioration does not proceed from larger sample sizes (Cochran, 1983). Internally, the issues of history, maturation, testing, instrumentation, statistical regression, selection, mortality, and various interactions of these variables dictate careful attention to a program evaluation's design (Campbell & Stanley, 1966). Likewise, interaction effects of selection biases and the dependent variable(s), the reactive effects of both pre-testing and the treatment setting, and the effects due to multiple treatments limit the generalizability of the program and may alter statistical testing. Note that the American Psychological Association's Task Force on Statistical Inference gives an interesting overview of the major areas for concern in any behavioral or social science research endeavor (Wilkinson, 1999). Even in the RCT's of clinical medicine, only rigorous attention to the details of design and methodology can avoid bias and the associated distortions of statistical results (Schulz, Chalmers, Hayes, & Altman, 1995).

Even if a particular program's design and methodology meets the above challenges, the issue of the meaning of a statistical significance test remains. Misinterpretations abound. To reiterate the corrections: the p-value is not the probability that the null hypothesis is true (Cohen, 1994); the p-value does not generally address replicability – see Abelson's (1995) discussion; and rejection of the null hypothesis does not affirm the theory being tested (Cohen, 1994). In the latter case, the null hypothesis of no difference between two populations invariably fails because, "...in the social sciences everything correlates with almost everything else" (Meehl, 1997), and a sufficiently large sample size will tease out the difference when couched in a statistical significance test (Thompson, 1993). The calculated p-value actually and merely expresses the probability (0 to 1.0) of observing a value of a particular test statistic as extreme (either as large or as small) or more extreme than the one observed, "given the sample size, and assuming that the sample was derived from a population in which the null hypothesis (H_0) is exactly true" (Thompson, 1996).

Most importantly, the p-value and the related NHST tell nothing about the value, magnitude, or importance of the substantive result. In program evaluation a statistically significant outcome may have no valid meaning for the client population who perceive little change in their status despite a reported positive outcome on some measure of knowledge, attitude, or behavior.

It should be noted that NHST receives a proper defense by Chow (1996), Abelson (1997), and Mulaik, Raju, and Harshman (1997). Chow, in particular, places NHST in the appropriate experimental and logical context, contexts that are so often lacking in program evaluation.

Other Approaches

Many of the NHST critics referenced above advocate a variety of alternatives including point estimation with confidence intervals, power analysis, effect sizes, and meta-analysis. These all offer an improvement on NHST, but they also fail to provide information about the absolute importance of the treatment effect or how many subjects actually reached any targeted improvement.

Confidence intervals present a range of values which contain a population parameter with a particular degree of probability, and they link closely with statistical significance testing (Gardner & Altman, 1986). Because of this link, they carry the same requirements for randomization and assumption satisfaction as NHST; they also fail "to settle issues raised about the inductive conclusion validity" for a study (Chow, 1996).

Statistical power, the ability to detect a particular effect size difference due to an intervention via statistical significance testing, also ties itself to the NHST requirements. Greater power or sensitivity derives from manipulations of the sample size, alpha level, statistical test, and effect size. However, Shaver (1993) wonders "What is the purpose of power analysis and the arbitrary manipulation of criteria in order to help ensure that the researcher will obtain a desired level of probability, when statistical significance has so little meaning?"

Effect sizes provide metrics that are independent of the sample size and can also be independent of the scale of measurement. "For a given dependent measure, effect size can be thought of simply as the difference between the means of the experimental versus control populations" (Lipsey, 1998). However, this absolute effect size has dependence on the scale of measurement. Standardizing the difference between the means allows the effect size to escape this dependency. The proportion of variance in the dependent variable that is explained by the independent variable offers another way to express effect size (Kellow, 1998). Various design and analysis decisions will have important consequences for the effect size observed (Posavac, 1998).

Unfortunately, the effect size, like NHST, can lead to misinterpretation. A rather arbitrary understanding of low, medium, and large effect sizes has developed with .2, .5, and .8 representing these values, respectively (Shaver, 1993). In many situations, relatively small effects may actually carry strong value, especially in cases "where there are processes by which individually tiny influences cumulate to produce meaningful outcomes" (Abelson, 1985), or in cases where either the independent variable undergoes only minimal manipulation or the dependent variable is difficult to influence (Prentice & Miller, 1992). At the opposite extreme, a large effect size does not necessarily translate into a meaningful outcome for the subjects of a program intervention (Jacobson & Truax, 1991; Lipsey & Wilson, 1993).

Effect sizes do enable the production of numerical averages for the combination of many studies through meta-analysis (Chow, 1988). Meta-analysis springs from appropriate effect size measures and their compatibility across studies, thus aggregating the strength of many evaluations. Meta-analysis has its own set of problems. Some arise from the reporting quality of the research database (Orwin & Cordray, 1985), particularly in regard to publication bias or the "file-drawer problem" where a number of studies, often those lacking statistical significance or sufficiently large effect size, cannot be located for inclusion in the analysis (Givens, Smith, & Tweedie, 1997). Others stem from a lack of independence across similar studies, confusion about the appropriate unit of analysis, confounding problems with weighted means, and misuse of tests of heterogeneity (Hall & Rosenthal, 1995). Rarely, meta-analyses relying on sets of small studies hold the potential to mislead practitioners, and a large RCT may be required to clarify understanding (Egger & Smith, 1995). Nevertheless, if employed with due diligence, meta-analysis reinforces the importance of replication for evaluative judgment.

Sustainable Clinical Improvement

Taking a more descriptive tack motivates several questions. What do we really need to know about the outcome of a program, and how can we report what we need to know given the available data? Statistical significance and effect sizes may or may not carry meaning for the clients in a particular program evaluation. This depends on what is important to the clients (with input from others) and whether an obtained valuable result continues beyond an immediate time frame as defined by the clients and evaluators. *Clinical significance* defines this valuable result or the minimal important difference (Guyatt, Juniper, Walter, Griffith, & Goldstein, 1998); in other words, clinical significance represents a treatment or intervention's "ability to meet standards of efficacy set by consumers, clinicians, and researchers" (Jacobson & Truax, 1991). This is the overriding substantive finding (Pedhazur, 1997), a finding whose definition rests increasingly with the patient according to some in the medical sciences (Guyatt et al., 1998).

Sustainable clinical improvement refers to client change that stabilizes at or above a previously agreed upon level of success for a previously agreed upon period of time. Some notion of eventual temporal stability should accompany the achievement of a targeted goal because an episodic distribution of success over time leaves a fleeting sense of accomplishment for many clients and because major fluctuations in benefit lead to underestimation or overestimation of the measure of efficacy (Laupacis, Sackett, & Roberts, 1988). Thus clients and evaluators must make two decisions up front: what level on a particular indicator represents a successful program outcome for a client and how long, within reason, must this level be maintained to acclaim that client truly improved with longevity. Thompson (1993) argues the need for these types of decisions and notes that "Statistics can be employed to evaluate the probability of an event. But importance is a question of human values, and math cannot be employed as an atavistic escape (a la Fromme's Escape from Freedom) from the existential human responsibility for making value judgements." The determination of value requires reasonable judgement.

The approach recommended here means that the criterion for value receives commitment before the program has started. Others follow a post hoc, statistical line of argument and allow the distributions of the treatment and comparison (control if truly experimental) groups to dictate the demarcation of program effectiveness. In addressing the issue of "what effect size is worth detecting?" Lipsey (1998) proposes Cohen's U3 measure (Cohen, 1977) and the binomial effect size display (BESD) of Rosenthal and Rubin (Rosenthal & Rubin, 1982) as possible candidates to answer the question. Cohen's U3 sets the mean of the control group as the success threshold while the BESD utilizes the grand median for the conjoint control and treatment distribution in the same role. Both assume normal distributions for either the control group or the control and treatment groups, respectively. With the normality assumption in place, Jacobson and Truax (Jacobson & Truax, 1991) develop a reliable change index utilizing the standard error of measurement to help determine if "real change" occurs. Although these methods have increasing value as a program nears the nature of a true experiment, we prefer controlling the criteria with reference to client, rather than statistical, input.

With sustainable clinical improvement established, the medical sciences provide a measure for quantifying a program's efficacy, the number needed to treat (NNT). Defined as the reciprocal of the absolute risk reduction (Laupacis et al., 1988), the NNT offers easy computation (Fig. 1) if the evaluator knows the total number of clients in the treatment (or active) and comparison groups and also the number who reached sustainable clinical improvement in each group. This information should always be available. Where a comparison group cannot be employed, the ratio formed for the comparison group (Ic/Tc) may be estimated from past experience or from similar populations. Note that the denominator in the NNT equation, the absolute risk reduction, is simply the difference between the event rate in the treatment group and the event rate in the comparison group, and this is a different type of expression for effect size (Hall & Rosenthal, 1995). An evaluator might take the approach of determining a critical effect size and asking if the mean of the treatment group exceeded this value when the comparison group, on average, did not. With the NNT, an evaluator asks how many of those in the treatment group, as a proportion, exceeded the critical effect size (defined by sustainable clinical improvement) than did those in the comparison group, as a proportion.

The lower the value for the NNT, the more efficacious is the program as measured on a particular indicator or dependent variable. A value of five means that five clients must receive the program's intervention or treatment for one client to reach sustainable clinical improvement. Obviously, a value of one indicates a perfect score; treating one client results in sustainable clinical improvement. Negative values favor the comparison group over the treatment group, and the program is actually impeding improvement in such cases. When the treatment and comparison groups have the same proportionate results, the NNT calculation involves division by zero yielding an undefined value, and this alerts the evaluator to the absence of efficacy. Very large values for the NNT make clear just how many clients must be served to achieve a single success, and the simplicity of the NNT promotes easy comparisons across different programs and for a single program with multiple applications over time.

Because the NNT comes from the experimentally oriented health science arena, the calculation of confidence intervals accompanies the statistic (Cook & Sackett, 1995). This is always desirable if the proper statistical warrant exists, and the intervals help to protect against embracing a strong, but not statistically significant, NNT that was generated from a small sample (Ware, Mosteller, Delgado, Donnelly, & Ingelfinger, 1992).

Using the simple head counting of the NNT avoids the parametric problems of means based statistics. The influence of outliers and the effects of averaging may make it seem that program efficacy has been achieved when, in terms of the number of clients actually achieving improvement, this is not the case. Often these numbers become lost in the good news that averaging can afford as Bracey (1999) demonstrates in the field of education.

Practical Significance

Many authors use the terms such as "practical significance" and "clinical significance" synonymously (Kirk, 1996; Lipsey & Wilson, 1993; Rosenthal, 1990). We propose distinguishing these terms by connecting *practical significance* to the efficiency of a program with reference to such variables as time, cost, contact hours, client satisfaction, and adverse effects. Clinical significance, of course, relates to a program's efficacy as discussed above.

The median time to reach sustainable clinical improvement for those clients achieving this goal provides one measure of practical significance. Likewise, the mean or median cost per program participant yields another such measure; contact hours are similarly computed. A Likert scale for overall client satisfaction, constructed with a rating of *one* being best, might render a median client rating as an indicator of efficiency. In a similar manner, evaluators might rate adverse effects and derive a median value.

Each of the above five measures has the property that the lower the value, the more efficient the program on that particular indicator. This generates ranking order for ordinal scales and relative values for interval or ratio comparisons across programs, and such measures as these tell us what we need to know about the practicality of programs. Combined indicators of efficacy and efficiency result from the products of the practicality measures and the NNT. Again, because both the practicality measures and the NNT have a lower/better connotation, the combined indicators express ranked values. While these types of indicators, along with the NNT itself, offer means for comparing programs of a similar nature or multiple applications of the same program over time or location, comparison of programs with different agendas would require additional factors for consideration, such as relevance, prevalence, severity, criticality, necessity, participation rates, etc.

Display of Information

An easy display of the above information for comparative purposes involves tabling the data with programs on the rows and indicators on the columns. The column order might lead with NNT and number of clients (or confidence intervals for the NNT if warranted) followed by the practicality indicators and then the combined indicators. A final column could contain some aggregate index of the combined indicators. Following Ehrenberg's (Bailar & Mosteller, 1988; Ehrenberg, 1977) rules gives such a table an organized and helpful arrangement.

Medical science offers another way to present the data in the form of a L'Abbe plot (L'Abbe, Detsky, & O'Rourke, 1987), designed to display the results of a meta-analysis in terms of outcome rates (Fig. 2). The event rate in the treatment group is plotted on the vertical axis while the event rate in the comparison group is plotted on the horizontal axis (these are the denominator terms in the NNT equation). Each point on the plot represents a single program or application of a program. A forty-five degree angle dotted line of equality separates the plot into two halves. If a point falls within the upper left half of the plot, then the treatment group fared better than the comparison group (a positive NNT). The exact position of the point tells how much better the treatment group did relative to the comparison group. A point in the lower right half of the plot represents a program where the comparison group bested the treatment group (a negative NNT). A program with equivocal results generates a point on the line of equality (an undefined NNT).

Following the example of the Bandolier web site (www.jr2.ox.ac.uk/bandolier/), we indicate the size of the program, in terms of the number of participating clients, via the size of a program's point or circle on the plot. In the absence of confidence intervals (always use confidence intervals if warranted), a program's sample size offers some relative indicator of confidence about the NNT when combined with a rigorous appraisal of design and methodological strength. Such an appraisal might be represented by assigning programs to various categories and coding these in color as shown in figure 3.

That is, categories A through C reflect declining quality in program selection, control, and application.

Color-coding has other uses, such as indicating the strength of treatment, value level for a practicality measure, subgroup assignment, year of program initiation, and level of attainment for the treatment group. In the context of the weight loss program example below, figures 4 and 5 illustrate some of these applications. Color-coding reinforces the impact of the L'Abbe plot by including important study characteristics with event rate effect size reporting and confidence intervals (or sample sizes) as suggested by Light, Singer, & Willett (1994).

Discussion

Jacobson and Truax (1991) use the example of a treatment for obesity to illustrate the difference between effect size and clinical significance. Suppose a quasi-experimental weight loss program for people averaging 300 lbs. showed that a statistically significant reduction (p=.05) had occurred with the average decline being two pounds. A 95% confidence interval about this two pound loss had a lower limit of one pound and an upper limit of three pounds. Power analysis indicated that the evaluation could have detected a half-pound weight difference with alpha set at .05 and a power of .90. Suppose further that a similar program, where the subjects had a standard deviation of five pounds, produced an average weight loss of four pounds for a relatively large effect size of .80. Also, a meta-analysis confirmed that this and similar programs yielded an average effect size of .65 with proportionate weight loss numbers.

How much meaningful value did the above programs render to their average client? Is a two pound or four pound weight loss really substantive to a 300 lb. person? Even if the mean weight loss were ten to twenty pounds, was this result due to a few clients with massive losses or many people with average losses? Did most clients maintain their loss for a reasonable period of time following cessation of intervention? How practical were these programs in relation to their efficacy? How do they compare in efficacy to other programs?

Utilizing sustainable clinical improvement and the indicators of practical significance produces answers to these questions in a consistent and straightforward manner where statistical inference is welcome, when warranted, but not necessary. For instance, if clients and evaluators defined sustainable clinical improvement as the loss of at least thirty pounds maintained for one year beyond cessation of the program's intervention, then an NNT of four would indicate that four people received treatment for one person to achieve success. The practicality measures and the combined indicators address efficiency/efficacy questions. L'Abbe plots provide the means for fairly easy interpretation of the results for subgroups in the same program, across replications of the same program, or in comparison to other programs. Figures 3 through 5 show some possible plots for such a weight loss program.

Shaver (1993) contends that "The question of interest is whether an effect size of a magnitude judged to be important has been consistently obtained across replications of adequate fidelity, not whether the result from a replication was statistically significant or whether the design had adequate power for a result to be statistically significant." Using sustainable clinical improvement forces the explicit definition of "a magnitude judged to be important" or of what Mosteller called the interocular difference (Scriven, 1993), the difference that hits us between the eyes. The NNT tells us how many attained such a difference in reference to what would have happened without a program's intervention. The practicality measures and combined indicators explicate the cost-benefit structure of a program; L'Abbe plots help us compare "replications of adequate fidelity."

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. Psychological Bulletin, 97(1), 129-133.

Abelson, R. P. (1995). Statistics as principled argument. Hillsdale, NJ: Lawrence Erlbaum Associates.

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 117-141). Mahwah, NJ: Lawrence Erlbaum Associates.

Bailar, J. C., III, & Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals. Annals of Internal Medicine, 108, 266-273.

Biskin, B. H. (1998). Comment on significance testing. Measurement and Evaluation in Counseling and Development, 31, 58-62.

Bracey, G. W. (1999). The forgotten 42%. Phi Delta Kappan, 80(9), 711-712.

Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasi-experimental designs for research. Chicago: Rand McNally and Company.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48(3), 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292.

Chow, S. L. (1988). Significance test or effect size? Psychological Bulletin, 103(1), 105-110.

Chow, S. L. (1996). Statistical significance: Rationale, validity and utility. London: Sage Publications.

Cochran, W. G. (1983). Planning and analysis of observational studies. New York: John Wiley & Sons.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. (rev. ed.). New York: Academic Press.

Cohen, J. (1994). The earth is round (p<.05). American Psychologist, 49(12), 997-1003.

Cook, R. J., & Sackett, D. L. (1995). The number needed to treat: A clinically useful measure of treatment effect. British Medical Journal, 310, 452-454.

Egger, M., & Smith, G. D. (1995). Misleading meta-analysis: Lessons from "an effective, safe, simple" intervention that wasn't. British Medical Journal, 310, 752-754.

Ehrenberg, A. S. C. (1977). Rudiments of numeracy. Journal of the Royal Statistical Society A, 140(3), 277-297.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. British Medical Journal, 292, 746-750.

Givens, G. H., Smith, D. D., & Tweedie, R. L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. Statistical Science, 12(4), 221-250.

Guyatt, G. H., Juniper, E. F., Walter, S. D., Griffith, L. E., & Goldstein, R. S. (1998). Interpreting treatment effects in randomised trials. British Medical Journal, 316, 690-693.

Hall, J. A., & Rosenthal, R. (1995). Interpreting and evaluating meta-analysis. Evaluation and the Health Professions, 18(4), 393-407.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). What if there were no significance tests? Mahwah, NJ: Lawrence Erlbaum Associates.

Hykle, J., Stevens, J. P., & Markle, G. (1993). Examining the statistical validity of studies comparing cooperative learning versus individualistic learning. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of consulting and clinical psychology, 59(1), 12-19.

Kellow, J. T. (1998). Beyond statistical significant tests: The importance of using other estimates of treatment effects to interpret evaluation results. American Journal of Evaluation, 19(1), 123-134.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56(5), 746-759.

L'Abbe, K. A., Detsky, A. S., & O'Rourke, K. (1987). Meta-analysis in clinical research. Annals of Internal Medicine, 107, 224-233.

Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. The New England Journal of Medicine, 318(26), 1728-1733.

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analyses. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 439-453). New York: Russell Sage Foundation.

Lipsey, M. W. (1998). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. J. Rog (Eds.), Handbook of applied social research methods (pp. 39-68). Thousand Oaks, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. American Psychologist, 48(12), 1181-1209.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46(4), 806-834.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 393-425). Mahwah, NJ: Lawrence Erlbaum Associates.

Moore, R. A., Gavaghan, D., Tramer, M. R., Collins, S. L., & McQuay, H. J. (1998). Size is everything - large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. Pain, 78, 209-216.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). The significance test controversy. Chicago: Aldine.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 65-115). Mahwah, NJ: Lawrence Erlbaum Associates.

Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. Psychological Bulletin, 97(1), 134-147.

Pedhazur, E. J. (1997). Multiple regression in behavioral research. Fort Worth: Harcourt Brace College Publishers.

Peers, J. (1978). 1001 logical laws, accurate axioms, profound principles, trusty truisms, homey homilies, colorful corollaries, quotable quotes, and rambunctious ruminations for all walks of life. New York: Fawcett Gold Medal.

Posavac, E. J. (1998). Toward more informative uses of statistics: Alternatives for program evaluators. Evaluation and Program Planning, 21, 243-254.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. Psychological Bulletin, 112(1), 160-164.

Rosenthal, R. (1990). How are we doing in soft psychology? American Psychologist, 45(6), 775-777.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology, 74(2), 166-169.

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias. Journal of the American Medical Association, 273(5), 408-412.

Scriven, M. (1993). Hard-won lessons in program evaluation. (Vol. 58). San Francisco: Jossey-Bass.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61(4), 293-316.

Stevens, J. (1996). Applied multivariate statistics for the social sciences. (3rd ed.). Mahwah, NJ: Lawrence Erlbaum associates.

The Standards of Reporting Trials Group. (1994). A proposal for structured reporting of randomized controlled trials. Journal of the American Medical Association, 272(24), 1926-1931.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61(4), 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Tukey, J. W., & Tukey, P. A. (1988). Computer graphics and exploratory data analysis: An introduction. In W. S. Cleveland (Ed.), The collected works of John W. Tukey, Graphics: 1965-1985 (Vol. 5, pp. 418-436). Pacific Grove, CA: Wadsworth & Brooks/Cole.

Ware, J. H., Mosteller, F., Delgado, F., Donnelly, C., & Ingelfinger, J. A. (1992). P values. In J. C. Bailar, III & F. Mosteller (Eds.), Medical uses of statistics (2nd ed., pp. 181-199). Boston: NEJM Books.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. Educational and PsychologicalMeasurement, 58(1), 21-37.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54(8), 594-604.

Figure 1

Number Needed To Treat (NNT)

	Active (Treatment)	Comparison
Total	Ta	Tc
Improved	Ia	Ic

$$NNT = \frac{1}{\underset{\swarrow}{(Ia/Ta)} - \underset{\searrow}{(Ic/Tc)}}$$

Treatment Event Rate Comparison Event Rate

The lower the better; 1 is perfect; negative values favor the comparison group; 1/0 is undefined - groups equal.

Figure 2

L'Abbe Plot

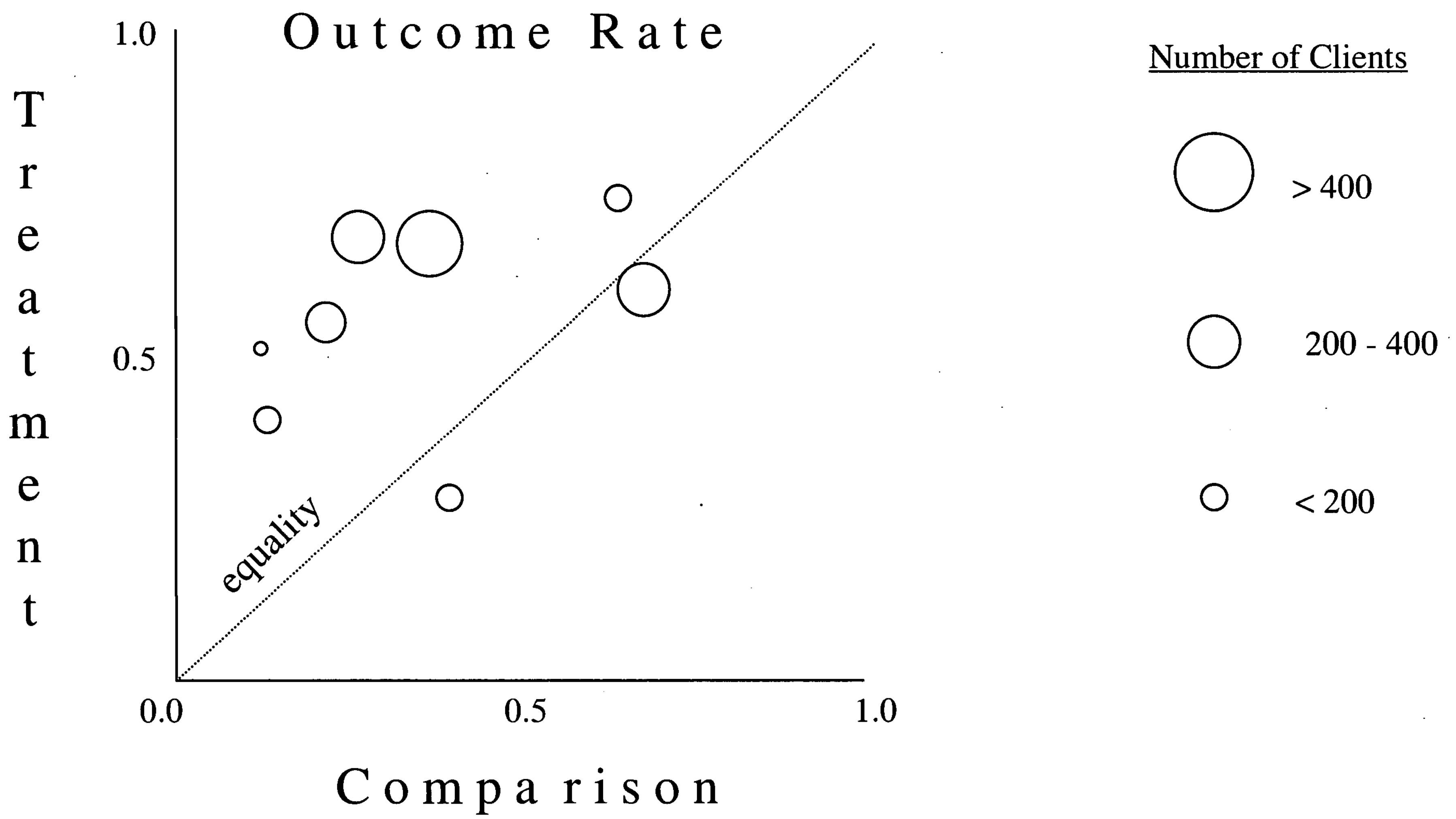


Figure 3

L'Abbe Plot
Comparison with Other Programs

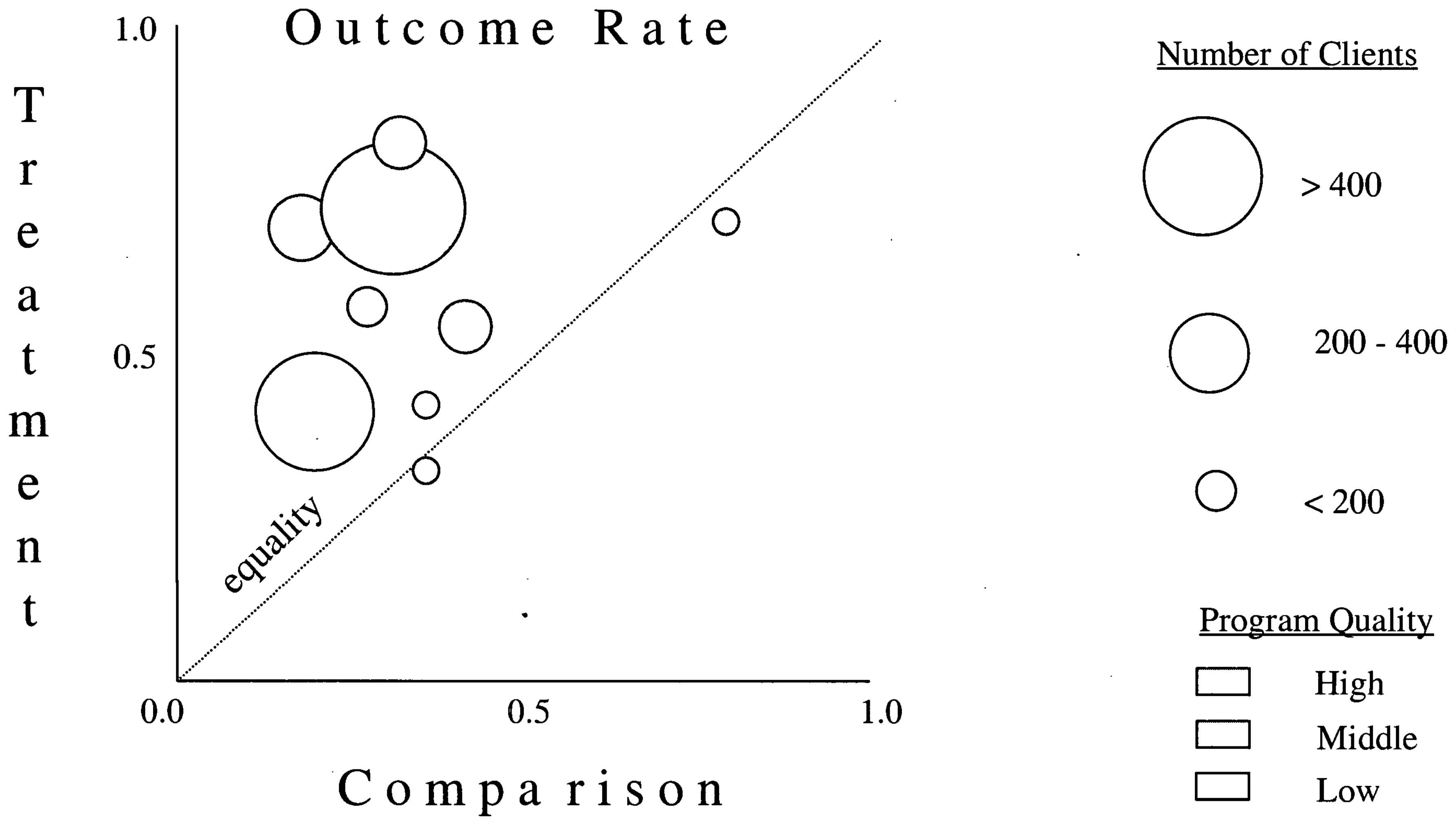


Figure 4

L'Abbe Plot
The Program Over Time

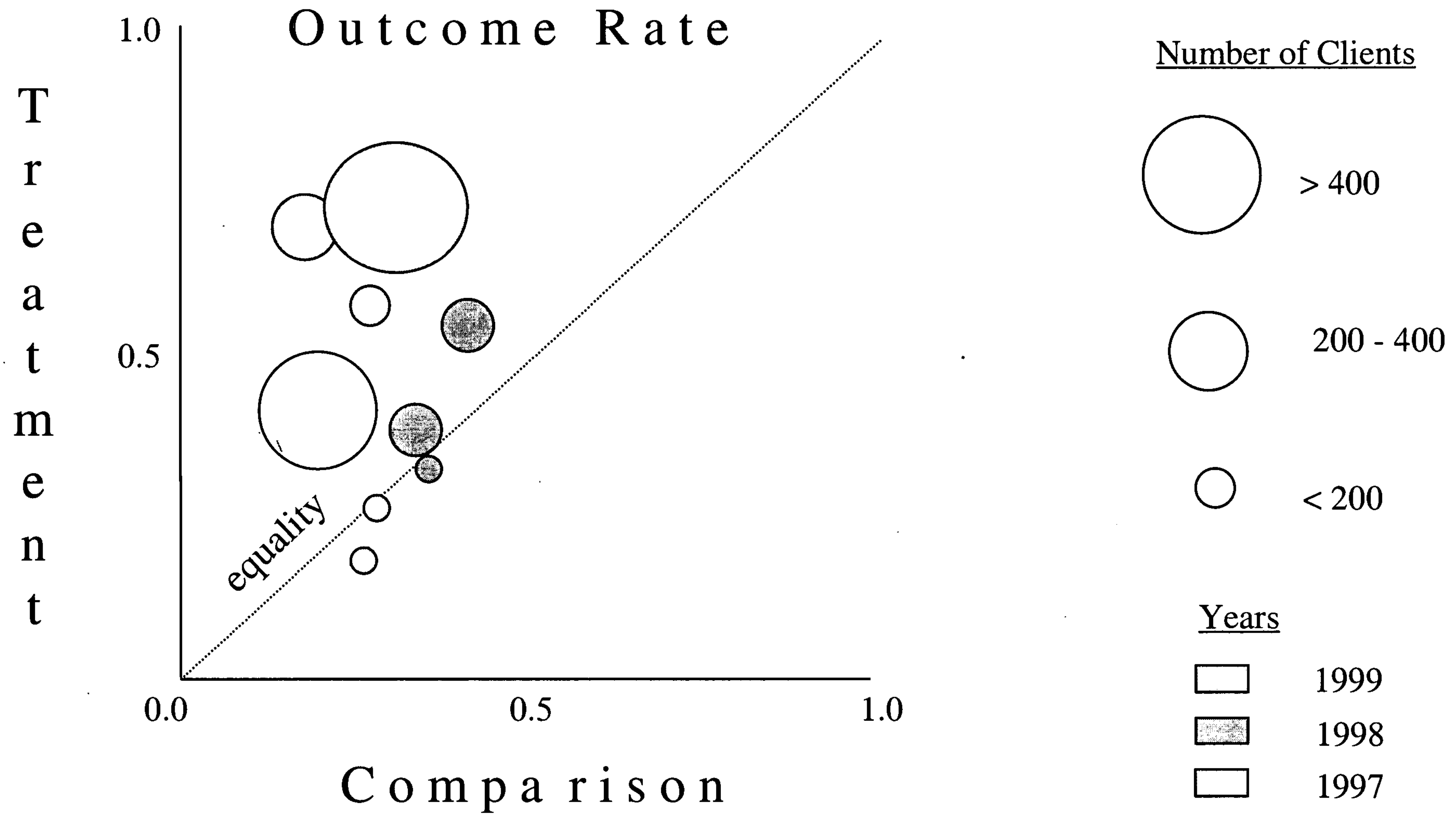
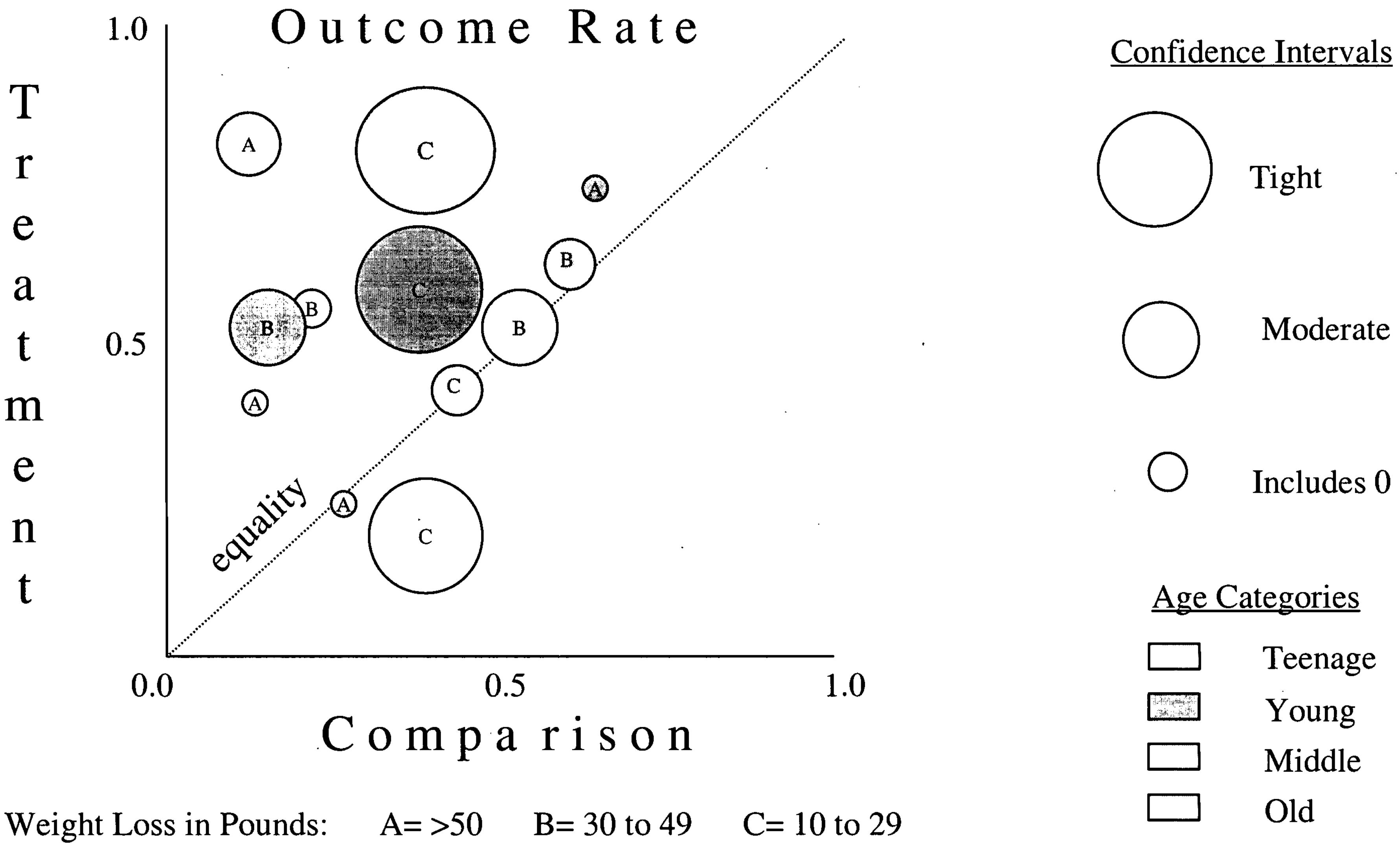


Figure 5

L'Abbe Plot
Subgroups and Sublevels



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: New Indicators for Program Evaluation	
Author(s): John C. Hanes and Michael Hail	
Corporate Source: University of North Carolina - Greensboro	Publication Date: 11/6/99

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
↑	↑	↑
☑	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic medie by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: John C. Hanes	Printed Name/Position/Title: John C. Hanes - Doctoral Student	
Organization/Address: University of North Carolina - Greensboro 1603 Pebble Drive Greensboro, N.C. 27410	Telephone: (336) 294-1875	FAX:
	E-Mail Address: jchanes@uncg.edu	Date: 12/2/99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.