Interpersonal expectations

Theory, research, and applications

Edited by

Peter David Blanck

Professor, University of Iowa College of Law, and Senior Fellow, the Annenberg Foundation Washington Program

CAMBRIDGE
UNIVERSITY PRESS

& Editions de la Maison des Sciences de l'Homme
Paris

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road,Oakleigh, Melbourne 3166, Australia
and
Editions de la Maison des Sciences de l'Homme
54 Boulevard Raspail, 75270 Paris, Cedex 06

First published 1993

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data
Interpersonal expectations : theory, research, and applications /
edited by Peter David Blanck.
p. cm. – (Studies in emotion and social interaction)
Includes index.
ISBN 0-521-41783-X (hard). – ISBN 0-521-42832-7 (pbk.)
1. Expectation (Psychology) 2. Interpersonal relations.
I. Blanck, Peter David, 1957– . II. Series.
BF323.E8I68 1993 92-36925
158'.2 – dc20 CIP

A catalog record for this book is available from the British Library.

ISBN 0-521-41783-X hardback
ISBN 0-521-42832-7 paperback
ISBN 2-7351-0492-3 hardback France only
ISBN 2-7351-0493-1 paperback France only

Contents

Introduction

1. Interpersonal expectations: Some antecedents and some consequences

ROBERT ROSENTHAL

It is a classic conception of progress that it is spiral in form. This volume, artfully conceived and orchestrated by Peter Blanck, lends support to that geometric conception. The spiral underlying this volume begins with the classic context-setting paper by Donald Campbell. That paper was presented at the first symposium on the social psychology of the psychological experiment, a 1959 symposium that dealt, in part, with the unintended effects of experimenters' expectations on the results of their experiments. It is now 1993, and this volume gives the current position of the moving spiral.

It is a heartening and enlightening experience to see more than 30 years of science go flashing by, beginning with Don Campbell's Janus-like classic that looks back to Francis Bacon's idols while it looks forward to the next generation of researchers. Campbell's classic, miraculously unpublished until this volume, seemed almost to have saved itself for this occasion. And some of the scholar-scientists of that next generation are also represented in this volume. All are young by my criteria, although ranging in career age from senior scholars of international renown to younger scholars in their first academic positions. But even the youngest scholars are already beginning to acquire that international renown.

One purpose of this introductory chapter is to review the history of some experiments designed to test the hypothesis of interpersonal expectancy effects. The earliest such experiments were designed to test this hypothesis in the context of the psychological experiment itself, with the experimenter serving as expecter and the research subject serving as expectee. What follows describes how this came about and then discusses some substantive and methodological consequences.

Experimenter expectancy effects and an unnecessary statistical analysis

As a graduate student at UCLA in the mid-1950s I was much taken with the work of two giants of personality theory, Freud and Murray. I was taken with Freud, as were so many others, for the richness and depth of his theory. I was taken with Murray, as were not enough others, for similar reasons but also because of Murray's brilliant way of inventing whatever tool was needed to further his inquiry. Thus, the Thematic Apperception Test (TAT) was invented simply as a tool to further his research, though it has become recognized as a major contribution in its own right. My dissertation was to depend on the work of both of these great theorists.

Sigmund Freud's projection

As a graduate student in clinical psychology I was (and still am) very much interested in projective techniques. Murray's TAT, Shneidman's Make a Picture Story Test, and, of course, the Rorschach were exciting methods for understanding people better. Shneidman, a brilliant researcher and clinician, was my first clinical supervisor during my Veterans Administration clinical internship. Bruno Klopfer, one of the all-time Rorschach greats, was the chair of my doctoral committee. It was natural, therefore, for me to be concerned with the defense mechanism of projection for the part it might play in the production of responses to projective stimuli.

Harry Murray's party game

Freud's defense mechanism of projection, the ascription to others of one's own states or traits (Freud, 1953; Rosenthal, 1956), is only one of the mechanisms that has been isolated as contributing to the process of producing responses to projective stimuli. Another mechanism is complementary apperceptive projection, that is, finding in another the reasons for one's own states or traits. It was this mechanism that Harry Murray investigated in his classic paper on "The Effect of Fear Upon Estimates of the Maliciousness of Other Personalities" (Murray, 1933). At his 11-year-old daughter's house party, Murray arranged a game called "Murder" that frightened delightfully the five party-going subjects. After the game, Murray found that the children perceived photo-

graphs as being more malicious than they did before the game. Murray's wonderfully direct and deceptively simple procedure of assessing projective processes by assessing changes in perceptions of photographs was the basic measuring device I adopted for my dissertation.

"An attempt at the experimental induction of the defense mechanism of projection"

With the foregoing as its almost unbearable title, my dissertation employed a total of 108 subjects: 36 college men, 36 college women, and 36 hospitalized patients with paranoid symptomatology. Each of these three groups was further divided into three subgroups receiving success, failure, or neutral experience on a task structured as, and simulating, a standardized test of intelligence. Before the subjects' experimental conditions were imposed, they were asked to rate the degree of success or failure of persons pictured in photographs. Immediately after the experimental manipulation, the subjects were asked to rate an equivalent set of photos on their degree of success or failure. The dependent variable was the magnitude of the difference scores between pre- and postratings of the photographs. It was hypothesized that the success condition would lead to the subsequent perception of other people as more successful, whereas the failure condition would lead to the subsequent perception of other people as having failed more, as measured by the pre- and postrating difference scores.

In an attack of studently compulsivity, an attack that greatly influenced my scholarly future, I did a statistical analysis that was extraneous to the main purpose of the dissertation. In this analysis I compared the mean *pre*-treatment ratings of the three experimental conditions. These means were: success = −1.52, neutral = −0.86, and failure = −1.02. The pre-treatment rating mean of the success condition was significantly lower than the mean of either of the other two conditions. It must be emphasized that these three treatment groups had not yet undergone their treatment; they were only destined to become the subjects of the three conditions. If the success group started out lower than the other groups, then, even if there were no differences among the three conditions in their post-treatment photo ratings, the success group would show the greatest gain, a result favoring one of my hypotheses, namely, that projection of the good could occur just as well as projection of the bad. Without my awareness, the cards had been stacked in favor of obtaining results supporting one of my hypotheses.

It should be noted that the success and failure groups' instructions had been identical during the pre-treatment rating phase of the experiment. (Instructions to the neutral group differed only in that no mention was made of the experimental task, since none was administered to this group.)

The problem, apparently, was that I knew for each subject which experimental treatment he or she would subsequently be administered. As I noted in 1956 with some dismay: "The implication is that in some subtle manner, perhaps by tone, or manner, or gestures, or general atmosphere, the experimenter, although formally testing the success and failure groups in an identical way, influenced the success subjects to make lower initial ratings and thus increase the experimenter's probability of verifying his hypothesis" (Rosenthal, 1956, p. 44). As a further check on the suspicion that success subjects had been treated differently, the conservatism extremeness of pre-treatment ratings of photos was analyzed. (The mean extremeness-of-rating scores were as follows: success = 3.92, neutral = 4.41, and failure = 4.42.) The success group rated photos significantly less extremely than did the other treatment groups. Whatever it was I did differently to those subjects whom I knew were destined for the success condition, it seemed to affect not only their mean level of rating but their style of rating as well.

The search for company

When I discussed these strange goings-on with some faculty members, they seemed not overly surprised. A not very reassuring response was "Oh, yes, we lose a few PhD dissertations now and then because of problems like that." There followed a frantic search of the literature for references to this phenomenon, which I then called *unconscious experimenter bias.* As far back as Ebbinghaus (1885), psychologists had been referring to something like this phenomenon, including such notables as Oskar Pfungst (1911), of Clever Hans fame, Ivan Pavlov (1929), and Saul Rosenzweig (1933). Unfortunately, none of these investigators (or even later ones) had explicitly designed and conducted an experiment to test the hypothesis of unconscious experimenter bias; that remained to be done.

There is something I want to add about the paper by Rosenzweig (1933), which appeared the same year as Harry Murray's paper (cited earlier) and, incidentally, the same year that I appeared. In my own several reviews of the literature (e.g., 1956, 1966), I had completely

missed the Rosenzweig paper. I believe it was my good friend, my long-time collaborator, and my scholarly tutor, Ralph Rosnow, who called my attention to Rosenzweig's extraordinarily insightful and prophetic paper. Not only did Rosenzweig anticipate the problem of unconscious experimenter bias, he also anticipated virtually the entire area now referred to as the *social psychology of the psychological experiment*. The Rosenzweig paper makes good reading even today, some 60 years later. There is a superb appreciation of the Rosenzweig paper in Ralph Rosnow's brilliant book about the methodology of social inquiry: *Paradigms in Transition* (Rosnow, 1981).

The production of company

If it was my unconscious experimenter bias that had led to the puzzling and disconcerting results of my dissertation, then presumably we could produce the phenomenon in our own laboratory, and with several experimenters rather than just one. Producing the phenomenon in this way would yield not only the scientific benefit of demonstrating an interesting and important concept, it would also yield the considerable personal benefit of showing that I was not alone in having unintentionally affected the results of my research by virtue of my bias or expectancy.

There followed a series of studies employing human subjects in which we found that when experimenters were led to expect certain research findings, they were more likely to obtain those findings. These studies were met with incredulity by many investigators who worked with human subjects. However, investigators who worked with animal subjects often nodded knowingly and told me that was the kind of phenomenon that encouraged them to work with animal subjects. In due course, then, we began to work with animal subjects and found that when experimenters were led to believe that they were working with maze-bright rats, the rats learned faster than did the rats randomly assigned to experimenters who had been led to believe that their rats were dull. That result surprised many psychologists who worked with animal subjects, but it would not have surprised Pavlov, Pfungst, or Bertrand Russell, who in 1927 had said: "Animals studied by Americans rush about frantically, with an incredible display of hustle and pep, and at last achieve the desired result by chance. Animals observed by Germans sit still and think, and at last evolve the solution out of their inner consciousness" (pp. 29–30).

Our experiments on the effects of investigators' expectancies on the behavior of their research subjects should be distinguished from the much older tradition of examining the effects of investigators' expectations, theories, or predilections on their observations or interpretations of nature. Examples of such effects have been summarized elsewhere (Rosenthal, 1966; see especially chapters 1 and 2 on observer effects and interpreter effects), and there is continuing lively interest in these topics (Gorman, 1986; Mahoney, 1989; Mitroff, 1974; Rudwick, 1986; Tweney, 1989).

Teacher expectation effects and an essential principal

If rats became brighter when expected to, then it should not be farfetched to think that children could become brighter when expected to by their teachers. Indeed, Kenneth Clark (1963) had for years been saying that teachers' expectations could be very important determinants of intellectual performance. Clark's ideas and our research should have sent us right into the schools to study teacher expectations, but that's not what happened.

What did happen was that after we had completed about a dozen studies of experimenter expectancy effects (we no longer used the term *unconscious experimenter bias*), I summarized our results in a paper for the *American Scientist* (Rosenthal, 1963). (As an aside, I should note that although this research had begun in 1958, and although there had been more than a dozen papers, none of them had been able to find their way into an American Psychological Association [APA] publication. I recall an especially "good news–bad news" type of day when a particular piece of work was simultaneously rejected by an APA journal and awarded the American Association for the Advancement of Science Socio-Psychological Prize for 1960. During these years of nonpublication, there were three "psychological sponsors" who provided enormous intellectual stimulation and personal encouragement: Donald T. Campbell, Harold B. Pepinsky, and Henry W. Riecken; I owe them all a great deal.)

I concluded this 1963 paper by wondering whether the same interpersonal expectancy effects found in psychological experimenters might not also be found in physicians, psychotherapists, employers, and teachers (subsequent research showed that indeed it could be found in all these practitioners). "When the master teacher tells his apprentice

that a pupil appears to be a slow learner, is this prophecy then self-fulfilled?" was the closing line of this paper (Rosenthal, 1963, p. 280).

Among the reprint requests for this paper was one from Lenore F. Jacobson, the principal of an elementary school in South San Francisco, California. I sent her a stack of unpublished papers and thought no more about it. On November 18, 1963, Lenore wrote me a letter telling of her interest in the problem of teacher expectations. She ended her letter with the following line: "If you ever 'graduate' to classroom children, please let me know whether I can be of assistance" (Jacobson, 1963).

On November 27, 1963, I accepted Lenore's offer of assistance and asked whether she would consider collaborating on a project to investigate teacher expectancy effects. A tentative experimental design was suggested in this letter as well.

On December 3, 1963, Lenore replied, mainly to discuss concerns over the ethical and organizational implications of creating false expectations for superior performance in teachers. If this problem could be solved, her school would be ideal, she felt, with children from primarily lower-class backgrounds. Lenore also suggested gently that I was "a bit naive" to think one could just *tell* teachers to expect some of their pupils to be "diamonds in the rough." We would have to administer some new test to the children, a test the teachers would not know.

Phone calls and letters followed, and in January 1964 a trip to South San Francisco to settle on a final design and to meet with the school district's administrators to obtain their approval. This approval was forthcoming because of the leadership of the school superintendent, Dr. Paul Nielsen. Approval for this research had already been obtained from Robert L. Hall, Program Director for Sociology and Social Psychology for the National Science Foundation, which had been supporting much of the early work on experimenter expectancy effects.

The Pygmalion experiment (Rosenthal & Jacobson, 1966, 1968)

All of the children in Lenore's school were administered a nonverbal test of intelligence, which was disguised as a test that would predict intellectual "blooming." The test was labeled the *Harvard Test of Inflected Acquisition*. There were 18 classrooms in the school, 3 at each of the six grade levels. Within each grade level, the three classrooms were composed of children with above-average, average, and below-average ability, respectively. Within each of the 18 classrooms, approximately 20%

of the children were chosen at random to form the experimental group. Each teacher was given the names of the children from his or her class who were in the experimental condition. The teacher was told that these children's scores on the Test of Inflected Acquisition indicated that they would show surprising gains in intellectual competence during the next 8 months of school. The only difference between the experimental group and the control group of children, then, was in the mind of the teacher.

At the end of the school year, 8 months later, all the children were retested with the same test of intelligence. Considering the school as a whole, the children from whom the teachers had been led to expect greater intellectual gain showed a significantly greater gain than did the children in the control group. The magnitude of this experimental effect was .30 standard deviation units, equivalent to a point biserial r of .15 (Cohen, 1988).

Some substantive consequences: Processes of social influence

Among the most interesting and important implications of the research on interpersonal expectancy effects are those for the study of subtle processes of unintended social influence. The early work in this area has been summarized in detail elsewhere (e.g., Rosenthal, 1966, 1969). When we look more particularly at the mediation of teacher expectancy effects, we find early summaries by Brophy and Good (1974), workers whose contributions to this area have been enormous, and by Rosenthal (1974). More recent summaries of this domain are by Brophy (1985) and by Harris and Rosenthal (1985). There is space here only to illustrate the types of research results that have been accumulating. A preliminary four-factor theory of the communication of expectancy effects suggests that teachers (and perhaps clinicians, supervisors, and employers) who have been led to expect superior performance from some of their pupils (clients, trainees, or employees) tend to treat these "special" persons differently than they treat the remaining less special persons in the following four ways (Rosenthal, 1971, 1973, 1974):

1. *Climate.* Teachers appear to create a warmer socioemotional climate for their special students. This warmth appears to be at least partially communicated by nonverbal cues.
2. *Feedback.* Teachers appear to give their special students more differentiated feedback, both verbal and nonverbal, as to how these students have been performing.

3. *Input.* Teachers appear to teach more material and more difficult material to their special students.
4. *Output.* Teachers appear to give their special students greater opportunities for responding. These opportunities are offered both verbally and nonverbally (e.g., giving a student more time in which to answer a teacher's question).

A recent simplification of the four-factor theory of the mediation of teacher expectation effects has been proposed (Rosenthal, 1989). This simplification, called the *affect/effort theory,* states that a change in the level of expectations held by a teacher for the intellectual performance of a student is translated into (1) a change in the affect shown by the teacher toward that student and, relatively independently, (2) a change in the degree of effort exerted by the teacher in teaching that student. Specifically, the more favorable the change in the level of expectation held by the teacher for a particular student, the more positive the affect shown toward that student and the greater the effort expended on behalf of that student. The increase in positive affect is presumed to be a reflection of increased liking for the student for any of several plausible reasons (Jussim, 1986). The increase in teaching effort is presumed to be a reflection of an increased belief on the part of the teacher that the student is capable of learning, so that the effort is worth it (Rosenthal & Jacobson, 1968; Swann & Snyder, 1980).

Some of the aspects of affect/effort theory currently under investigation with Nalini Ambady have very exciting implications. For example, we have been able to predict student ratings of a college instructor's effectiveness over the course of an entire semester from an examination of a 30-second slice of teaching behavior in which we have access only to the silent videotape or to the tone of voice (not the content) in which the instructors are communicating with their students. These predictive correlations, often in the range of .6 to .7, have been replicated in high school settings and fit very well with the results of many other studies of "thin slices" of nonverbal behavior summarized meta-analytically (Ambady & Rosenthal, 1992, 1993).

Similarly, work with Sarah Hechtman has shown the potential for affect/effort theory to help explain the traditional sex differences in cognitive functioning. We have found that teachers teaching verbal material to males and quantitative material to females (the so-called sex-inappropriate materials) showed greater hostility to their students in the nonverbal channels (video-only) than did teachers teaching the so-

called sex-appropriate materials to these same students. These bias effects were smaller for female than for male teachers, and they were smaller for more androgynous than for more sex-typed teachers (Hechtman & Rosenthal, 1991).

Some methodological consequences for a better understanding of replication

Unfriendly reactions to the research on interpersonal expectancy effects and claims of failures to replicate the effects led me to examine closely and, no doubt, defensively the concept of replication in behavioral research (Rosenthal, 1966, 1990).

There is a long tradition in psychology of our urging one another to replicate each other's research. But, although we have been very good at calling for replication, we have not been very good at deciding when a replication has been successful. The issue we now address is: When shall a study be deemed successfully replicated?

Successful replication is ordinarily taken to mean that a null hypothesis that has been rejected at Time 1 is rejected again, and with the same direction of outcome, on the basis of a new study at Time 2. We have a failure to replicate when one study was significant and the other was not. Let us examine more closely a specific example of such a failure to replicate.

Pseudo-failures to replicate

The saga of Smith and Jones. Smith has published the results of an experiment in which a certain treatment procedure was predicted to improve performance. She reported results significant at $p < .05$ in the predicted direction. Jones publishes a rebuttal to Smith, claiming a failure to replicate. In situations of that sort, it is often the case that, although Smith's results were more significant than Jones's, the studies were in quite good agreement as to their estimated sizes of effect, as defined either by Cohen's d $[(\text{mean}_1 - \text{mean}_2)/\sigma]$ or by r, the correlation between group membership and performance score (Cohen, 1988; Rosenthal, 1991). Thus, studies labeled as *failures to replicate* often turn out to provide strong evidence for the replicability of the claimed effect.

On the odds against replicating significant results. A related error often found in the behavioral and social sciences is the implicit assumption

that if an effect is real, we should therefore expect it to be found significant again upon replication. Nothing could be further from the truth.

Suppose that there is in nature a real effect with a true magnitude of $d = .50$ [i.e., $(\text{mean}_1 - \text{mean}_2)/\sigma = .50$ σ units] or, equivalently, $r = .24$ (a difference in success rate of 62% versus 38%). Then suppose that an investigator studies this effect with an N of 64 subjects or so, giving the researcher a level of statistical power of .50, a very common level of power for behavioral researchers of the last 30 years (Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Even though a d of .50 or an r of .24 can reflect a very important effect, there is only one chance in four that both the original investigator and a replicator will get results significant at the .05 level. If there were two replications of the original study, there would be only one chance in eight that all three studies would be significant, even though we know that the effect in nature is real and important.

Contrasting views of replication

The traditional, not very useful, view of replication has two primary characteristics:

1. It focuses on significance level as the relevant summary statistic of a study, and
2. it evaluates whether replication has been successful in a dichotomous fashion. For example, replications are successful if both or neither $p < .05$ and they are unsuccessful if one $p < .05$ and the other $p > .05$. Psychologists' reliance on a dichotomous decision procedure accompanied by an untenable discontinuity of credibility in results varying in p levels has been well documented (Nelson, Rosenthal, & Rosnow, 1986; Rosenthal & Gaito, 1963, 1964).

The newer, more useful views of replication success have two primary characteristics:

1. A focus on effect size as the more important summary statistic of a study, with a relatively minor interest in the statistical significance level, and
2. An evaluation of whether replication has been successful is made in a continuous fashion. For example, two studies are not said to be successful or unsuccessful replicates of each other; rather, the degree of failure to replicate is specified.

Some metrics of the success of replication

Differences between effect sizes. Once we adopt a view of the success of replication as a function of the similarity of effect sizes obtained, we can become more precise in our assessments of the success of replication. Replication success can be indexed by the difference between the effect sizes obtained in the original study and in the replication. For example, we could employ the differences in Cohen's *d*'s or the effect size *r*'s obtained, or we could employ Cohen's *q*, which is the difference between *r*'s that have been first transformed to Fisher's *Z*'s. Fisher's *Z* metric is distributed nearly normally and can thus be used in setting confidence intervals and testing hypotheses about *r*'s, whereas *r*'s distribution is skewed, and the more so as the population value of *r* moves further from zero. Cohen's *q* is especially useful for testing the significance of difference between two obtained effect size *r*'s (Rosenthal, 1991; Rosenthal & Rubin, 1982; Snedecor & Cochran, 1989). When there are more than two effect size *r*'s to be evaluated for their variability (i.e., heterogeneity), we can simply compute the standard deviation (*S*) among the *r*'s or their Fisher *Z* equivalents. If a test of significance of heterogeneity of these Fisher *Z*'s is desired, a simple χ^2 test of heterogeneity is readily available (Hedges, 1982; Rosenthal & Rubin, 1982).

Meta-analytic metrics. As the number of replications for a given research question grows, a full assessment of the success of the replicational effort requires the application of meta-analytic procedures. An informative summary of the meta-analysis might be the stem-and-leaf display of the effect sizes found in the meta-analysis (Tukey, 1977). A more compact summary of the effect sizes might be Tukey's (1977) box plot, which gives the highest and lowest obtained effect sizes, along with those found at the 25th, 50th, and 75th percentiles. For single index values of the consistency of the effect sizes, one could employ (1) the range of effect sizes found between the 75th (Q_3) and 25th (Q_1) percentiles, (2) some standard fraction of that range (e.g., one-half or three-quarters), (3) *S*, the standard deviation of the effect sizes, or (4) *SE*, the standard error of the effect sizes.

As a slightly more complex index of the stability, replicability, or clarity of the average effect size found in the set of replicates, one could employ the mean effect size divided either by its standard error ($\sqrt{k}$, where *k* is the total number of replicates) or simply by *S*. The latter index of mean effect size divided by its standard deviation (*S*) is the

reciprocal of the coefficient of variation or a kind of coefficient of robustness.

The coefficient of robustness of replication. Although the standard error of the mean effect size, along with confidence intervals placed around the mean effect size, are of great value (Rosenthal & Rubin, 1978), it will sometimes be useful to employ a robustness coefficient that does not increase simply as a function of the increasing number of replications. Thus, if we want to compare two research areas for their robustness, adjusting for the difference in number of replications in each research area, we may prefer the robustness coefficient defined as the reciprocal of the coefficient of variation.

The utility of this coefficient is based on two ideas – first, that replication success, clarity, or robustness depends on the homogeneity of the obtained effect size, and second, that it depends also on the clarity of the directionality of the result. Thus, a set of replications grows in robustness as the variance of the effect sizes decreases and as the distance of the mean effect size from zero increases. Incidentally, the mean may be weighted, unweighted, or trimmed (Tukey, 1977). Indeed, it need not be the mean at all but any measure of location or central tendency (e.g., the median).

What should be reported?

Effect sizes and significance tests. If we are to take seriously our newer view of the meaning of the success of replications, what should be reported by authors of papers seen to be replications of earlier studies? Clearly, reporting the results of tests of significance will not be sufficient. The effect size of the replication and of the original study must be reported. It is not crucial which particular effect size is employed, but the same effect size should be reported for the replication and the original study. Complete discussions of various effect sizes and when they are useful are available from Cohen (1988) and elsewhere (e.g., Rosenthal, 1991, in press). If the original study and its replication are reported in different effect size units, these can usually be translated to one another (Cohen, 1988; Rosenthal, 1991; Rosenthal & Rosnow, 1991; Rosenthal & Rubin, 1989).

Power. Especially if the results of either the original study or its replication were not significant, the statistical power at which the test