Psychological Methods
1996, Vol. 1, No. 3, 278–292

1082-989X/96/$3.00

Directional Decisions for Two-Tailed Tests: Power, Error Rates, and Sample Size

Les Leventhal and Cam-Loi Huynh
University of Manitoba

One imposing directional decisions on nondirectional tests will overestimate power, underestimate sample size, and ignore the risk of Type III error (getting the direction wrong) if traditional calculations—those applying to nondirectional decisions—are used. Usually trivial with the z test, the errors might be important where α is large and effect size is small or with tests using other distributions. One can avoid the errors by using calculations that apply to directional decisions or by using a *directional* two-tailed test at the outset, a conceptually simpler solution. With a revised concept of power, this article shows calculations for the test; explains how to find its power, Type III error risk, and sample size in statistical tables for traditional tests; compares it to conventional one- and two-tailed tests and to one- and two-sided confidence intervals; and concludes that when a significance test is planned it is the best choice for most purposes.

Researchers frequently investigate the direction rather than the size of a relationship, for example, when testing a theory that can predict only the direction of an effect or when investigating "Which is more?" or "Which is better?" These are some of the fundamental questions in social science. But evaluating direction with traditional significance tests has problems. One-tailed tests provide for deciding whether or not an effect falls in the direction predicted by the researcher but do not provide for deciding, whatever the results, that the effect falls in the opposite direction. Conventional two-tailed tests are nondirectional: They evaluate nondirectional statistical hypotheses and do not provide for directional decisions. Nevertheless, a common solution is to conduct a nondirectional two-tailed test and, after finding significance, to inspect data visually to make a directional decision. An example would be a significant mean difference test in which one inspects the sample means to decide the direction of the difference between the population means. Many statistics texts and statistical writers use this solution (e.g., Bakan, 1966, p. 431; Cohen, 1965, p. 108; Hildebrand, 1986, pp. 319–320; Keppel, 1991, p. 123; Kiess, 1996, pp. 178, 228; May, Masson, & Hunter, 1990, pp. 243–245, 268–270; McClave & Sincich, 1995, p. 323, 385; Minium, King, & Bear, 1993, p. 284; Pagano, 1994, pp. 231, 234; Spatz & Johnston, 1989, pp. 156–157). Aside from the obvious problem that the decision reached has little to do with the statistical hypotheses tested, this solution has difficulties. First, traditional power and sample size calculations for the nondirectional test apply to nondirectional decisions, not to directional decisions. But power can differ for the two decisions, and therefore so can sample size needed for a given power. A calculational example below illustrates a two-tailed mean difference test that requires 55 subjects if planning a nondirectional decision but requires 59 subjects if planning a directional decision. The second difficulty with imposing a directional decision on a nondirectional test is that no provision is made for calculating the risk of getting the direction wrong, often referred to as *Type III error*.

One planning a directional decision with a nondirectional two-tailed test can avoid these problems by

Les Leventhal and Cam-Loi Huynh, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada.

This work was supported in part by the Research Grants Committee of the University of Manitoba. We are grateful to James B. Nickels and Rhonda K. Kowalchuk, both of the Department of Psychology, University of Manitoba, for comments on an earlier version of this article. Any errors remaining are our responsibility.

Correspondence concerning this article should be addressed to Les Leventhal, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2.

using the methods described below to calculate power, error rates, and sample size for directional decisions. A better alternative is to conduct a *directional* two-tailed test from the outset, a conceptually simpler solution. Either way, the calculations apply to directional decisions. The purpose of this article is to describe how to conduct a directional two-tailed test; to describe how to calculate error rates, power, and sample size for directional decisions; to discuss why the calculations presuppose a revised definition of power; and to compare the directional test to conventional one- and two-tailed tests and to one- and two-sided confidence intervals.

Conducting a Directional Two-Tailed Test

Hodges and Lehmann (1954) and Kaiser (1960) discussed the directional two-tailed test but without sufficient detail. Ferguson and Takane (1989, p. 184) referred to Kaiser's article but incorrectly explained the significance level of the test. Hand, McCarter, and Hand (1985) described how to conduct the test by using confidence intervals but were vague on how to compute power, error rates, and sample size. We describe how to conduct the test with a significance testing approach.

There are two equivalent ways of understanding the directional two-tailed test (Kaiser, 1960). One way is to view it as two simultaneous one-tailed tests predicting opposite directions. For example, when testing the difference between sample means, one one-tailed test assesses

$$H_0\text{: } \mu_1 - \mu_2 \leqslant 0^{1}$$
$$H_A\text{: } \mu_1 - \mu_2 > 0$$

and the other assesses

$$H_0\text{: } \mu_1 - \mu_2 \geqslant 0$$
$$H_A\text{: } \mu_1 - \mu_2 < 0.$$

If the directional two-tailed test uses overall significance level α, then each one-tailed test would normally use 0.5α. (The one-tailed test significance levels must sum to α but need not be equal; see Shaffer, 1972.) If a one-tailed test rejects its H_0 in favor of H_A at 0.5α, then one decides at α that the population means differ in the direction of the accepted H_A. With overall α of 0.05 as an example, if either (a) Sample Mean 1 is sufficiently larger than Sample Mean 2 so that the probability is 0.025 or less when $\mu_1 - \mu_2 = 0$, allowing one-tailed rejection of H_0: $\mu_1 - \mu_2 \leqslant 0$ in favor of H_A: $\mu_1 - \mu_2 > 0$, or (b) Sample Mean 1 is sufficiently smaller than Sample Mean 2 so that the probability is 0.025 or less when $\mu_1 - \mu_2 = 0$, allowing one-tailed rejection of H_0: $\mu_1 - \mu_2 \geqslant 0$ in favor of H_A: $\mu_1 - \mu_2 < 0$, then one decides at the 0.05 level in a two-tailed test that the population means differ in the direction of the accepted H_A. If neither Case a nor b occurs, then one does not reject either H_0. According to Shaffer (1972), "the overall Type I probability error is . . . the probability of rejecting at least one of the [null] hypotheses, given that both are true, [and this] is the sum of the maximum Type I error probabilities, that is, the sum of the sizes of the two tests" (p. 196; see also Hodges & Lehmann, 1954, p. 264).

The other way to understand the directional two-tailed test is to view it as a single test assessing three statistical hypotheses: H_1, H_2, and H_3. When testing the difference between two sample means, the hypotheses are

$$H_1\text{: } \mu_1 - \mu_2 < 0$$
$$H_2\text{: } \mu_1 - \mu_2 = 0 \text{ (null hypothesis)}$$
$$H_3\text{: } \mu_1 - \mu_2 > 0.$$

H_2 is the null hypothesis, that is, the hypothesis generating the sampling distribution containing the rejection regions. Again, using overall significance level α of 0.05 as an example, if either (a) Sample Mean 1 is sufficiently larger than Sample Mean 2 so that the probability is 0.025 or less when H_2 is true, allowing rejection of H_2 in favor of H_3, or (b) Sample Mean 1 is sufficiently smaller than Sample Mean 2 so that the probability is 0.025 or less when H_2 is true, allowing rejection of H_2 in favor of H_1, then one decides at the

[1] Many texts would use inexact H_0: $\mu_1 - \mu_2 \leqslant 0$, and many texts would use exact H_0: $\mu_1 - \mu_2 = 0$ for this one-tailed test's null hypothesis. Either is acceptable (Kaiser, 1960). For the inexact H_0, the point value $\mu_1 - \mu_2 = 0$ falls at the edge of the null values and generates the sampling distribution for the test. When evaluating H_0: $\mu_1 - \mu_2 \leqslant 0$ against H_A: $\mu_1 - \mu_2 > 0$, data that reject $\mu_1 - \mu_2 = 0$ at $\alpha = 0.05$ will reject at $\alpha < 0.05$ any H_0 for which $\mu_1 - \mu_2 < 0$ (see Hays, 1981, pp. 232, 254–257; Kirk, 1984, p. 250). With the inexact null, α is usually understood to mean the maximum probability of Type I error (e.g., Hodges & Lehmann, 1954, p. 264; Kaiser, 1960; Kirk, 1982, p. 31; Shaffer, 1972, p. 195). Although inexact nulls better express the hypotheses under test, it makes no practical difference whether exact or inexact nulls appear in the one-tailed tests that constitute the directional two-tailed test.

0.05 level in a two-tailed test that the difference between the population means falls in the direction of the accepted hypothesis. (The probabilities in Cases a and b must sum to α but need not be equal; see Braver, 1975, Hick, 1952, and Nosanchuk, 1978, who discussed splitting α unequally between the tails of a conventional two-tailed test, a discussion that also applies to the directional test.) The overall α is 0.05 because data satisfying Cases a and b constitute the extreme values that, altogether, occur 5% of the time when H_2 is true. So the risk of rejecting a true H_2 is 0.05. If neither Case a nor b occurs, then there is no reason to reject H_2. For convenience, future discussion adopts this three-hypothesis interpretation of the directional test, and we often call it a *three-choice* test.

Since the three-choice test and the conventional nondirectional test are two-tailed, each test has two critical values. Moreover, given the same α, the tests use the same critical values and make the same decision about the null hypothesis. After null rejection, however, the tests accept different alternative hypotheses. The conventional test accepts a nondirectional alternative, and the three-choice test accepts one of the directional alternatives by visually determining on which side of the null hypothesis the obtained data fall. Although the three-choice test and the conventional test use the same critical values for null rejection, they can differ in error rates, power, and sample size. We first discuss error rates and power.

Error Rates and Power

The traditional definition of power—probability of rejecting a false null—is usually discussed in the context of a significance test that provides for two exhaustive and mutually exclusive states of nature (H_0 true or H_A true) and two exhaustive and mutually exclusive decisions (accept H_0 or accept H_A; we use accept or retain H_0 rather than not reject H_0 in order to simplify language). This produces the familiar fourfold table shown in Figure 1, which shows correct and incorrect decisions and their conditional probabilities. Note that the power decision of rejecting a false null entails accepting a correct alternative (H_A). This will not be the case with the three-choice test, discussed presently. First, we discuss error rates for a three-choice test.

Error Rates

A three-choice test provides for three exhaustive and mutually exclusive states of nature (H_1 true, H_2 true, or H_3 true) and three exhaustive and mutually exclusive decisions (accept H_1, accept H_2, or accept H_3), producing the ninefold table shown in Figure 2. Here, H_2 is the null hypothesis, and H_1 and H_3 are directional alternatives. Like the fourfold table in Figure 1, (a) cells in the column under any state of nature are exhaustive and mutually exclusive, (b) each cell has a conditional probability understood as the probability of the decision in the cell's row given the state of nature, reduced to a point value, in the cell's column (H_2 is already a point value), (c) probabilities for the cells in any column must sum to 1, and (d) the probability for a cell in a column can be obtained by subtracting the probabilities of the other cells in the column from 1. We explain the unexpected power entries in the four corners of the table in the section on power.

		NATURE	
		H_0 true	H_A true
DECISION ABOUT NATURE	Accept H_0	Correct $1-\alpha$	Type 2 error β
	Accept H_A	Type 1 error α	Correct $1-\beta$ Power

Figure 1. Fourfold table showing two decisions (their correctness and conditional probability) for each of two states of nature.

A Type I error, having conditional probability α, is the rejection of a true null hypothesis. This error is possible only when H_2 is true. Two cells in the middle column make different versions of this error: the cell marked α_{12}, in which one accepts H_1 when H_2 is true, and the cell marked α_{32}, in which one accepts H_3 when H_2 is true. (The first subscript for α indicates the hypothesis that is accepted, and the second indicates the hypothesis that is true.) When H_2 is true, either cell can occur, and the probability of one or the other occurring under H_2 is the sum of their separate probabilities: $\alpha_{12} + \alpha_{32}$. This sum is the overall α for a three-choice test.

A Type II error, having conditional probability β, is the acceptance of a false null hypothesis. This error is possible only when the null hypothesis is false, that is, when H_1 or H_3 is true. Two cells in the middle row, β_{21} (accept H_2 when H_1 is true) and β_{23} (accept H_2

		NATURE		
		H_1 true	H_2 true	H_3 true
DECISION ABOUT NATURE	Accept H_1	Correct $1-\beta_{21}-\gamma_{31}$ Power_{11}	Type 1 error α_{12}	Type 3 error γ_{13} Power_{13}
	Accept H_2	Type 2 error β_{21}	Correct $1-\alpha_{12}-\alpha_{32}$	Type 2 error β_{23}
	Accept H_3	Type 3 error γ_{31} Power_{31}	Type 1 error α_{32}	Correct $1-\beta_{23}-\gamma_{13}$ Power_{33}

Figure 2. Ninefold table showing three decisions (their correctness and conditional probability) for each of three states of nature. This table uses the traditional definition of power.

when H_3 is true), make different versions of this error. When a given alternative hypothesis is true, only one of these cells can occur, and the probability of the cell occurring under that alternative is the probability of Type II error for the three-choice test. Hence, the probability of Type II error is β_{21} when H_1 is true and β_{23} when H_3 is true. One can compute the probability of Type II error only when a specific parameter value is stated by H_1 or H_3. Since both the directional and nondirectional two-tailed tests use the same null hypothesis and the same critical values for null rejection, both tests make the same decision about the null hypothesis whether the null is true or false. So the tests have the same risk of accepting a false null hypothesis and therefore the same risk of Type II error.

A Type III error, having conditional probability γ, is the acceptance of one directional alternative when the other is true: γ is the probability of getting the direction wrong. (The definition of Type III error varies from author to author; e.g., compare Hopkins, 1973, Kaiser, 1960, and Mosteller, 1948). This error is possible only when H_1 or H_3 is true. Two cells, γ_{31} (accept H_3 when H_1 is true) and γ_{13} (accept H_1 when H_3 is true), make different versions of this error. When a given alternative hypothesis is true, only one of these cells can occur and the probability of the cell occurring under that alternative is the probability of Type III error for the three-choice test. For example, when H_3 is true, the only Type III error possible is the acceptance of H_1; hence, γ_{13} is the probability of Type III error when H_3 is true. One can compute the probability of Type III error only when a specific parameter value is stated by H_1 or H_3. The nondirectional two-tailed test does not provide for a directional decision and, hence, cannot make a Type III error. A one-tailed test can make a Type III error by accepting directional alternative H_A when the truth falls in the opposite direction (see Shaffer, 1972).

Power

Traditional Definition

Power is traditionally defined as the (conditional) probability of rejecting a false null hypothesis. When the null hypothesis, H_2, is true, power is undefined. One can compute a probability for power only when a specific parameter value is stated by H_1 or H_3. The four corner cells of Figure 2, marked *power,* make different versions of the power decision. This is clearly the case with the corners in the upper left (accept H_1 when H_1 is true) and lower right (accept H_3 when H_3 is true). But this is also the case with the corners in the upper right (accept H_1 when H_3 is true) and lower left (accept H_3 when H_1 is true). For example, while accepting H_3 when H_1 is true is a Type III error, it is also a correct decision in the power sense of rejecting a false null: The null must be false when H_1 is true, and one must reject the null to accept H_3. Accordingly, this cell is marked both γ_{31} and power_{31}: γ_{31} because it represents the Type III error of accepting H_3 when H_1 is true, and power_{31} because accepting H_3 when H_1 is true also counts as the correct power decision of rejecting a false null. This unexpected state of affairs results from the fact that, in a three-choice test, the power decision of rejecting a false null hypothesis does not entail accepting a correct alternative. Accordingly, in the three-choice test, the traditional definition of power produces the uncomfortable result that power can include incorrect decisions.

The traditional definition of power also results in the three-choice test having the same power as the nondirectional test. The three-choice test and the nondirectional test use the same null hypothesis and the same critical values for null rejection. Hence, both tests make the same decision about the null hypothesis whether the null is true or false. So both tests have the same probability of rejecting a false null hypothesis and therefore have the same power.

Revised Definition

A different definition of power, one based on that of Mosteller (1948), will be useful for evaluating directional decisions, namely, the (conditional) probability of rejecting a false null hypothesis in favor of a true alternative. The traditional definition—probability of rejecting a false null—fails to specify which hypothesis to accept after null rejection. The change in definition has no effect on conventional one- or two-tailed tests. Since conventional tests have only two hypotheses, rejecting a false null hypothesis necessarily implies accepting a true alternative. For the three-choice test, however, the new definition excludes from power the Type III errors in the upper right and lower left corners of Figure 2, producing a new ninefold table shown in Figure 3 (cf. Kaiser, 1960).

Which definition of power is more useful? An investigator planning a directional decision with a two-tailed test will probably find it more useful to calculate the probability of rejecting a false null hypothesis in favor of a correct directional decision (revised definition) than of rejecting a false null hypothesis in favor of a directional decision that may or may not be correct (traditional definition). A literature reviewer conducting a power analysis to interpret nonsignificant findings from a nondirectional test in a published study would probably find it more useful to calculate power for reaching a correct directional decision (revised definition) if the reviewer were interested in direction.

With the revised definition, the three-choice directional test is less powerful than is the nondirectional test. This can be seen by comparing the ninefold tables (Figures 2 and 3). With the traditional definition of power in Figure 2, the three-choice test's power is

$$\text{power} = 1 - \beta \qquad (1)$$

for a given state of nature. For example, when H_3 is true, power under H_3 equals power$_{13}$ + power$_{33}$ = 1 − β_{23}. When H_1 is true, power under H_1 is power$_{11}$ + power$_{31}$ = 1 − β_{21}. With the revised definition in Figure 3, γ no longer counts as power, and power is therefore reduced by that amount. Accordingly, with the revised definition, the three-choice test's power is

$$\text{power} = 1 - \beta - \gamma \qquad (2)$$

for a given state of nature. For example, when H_3 is true, the revised definition reduces power by an amount equal to γ_{13}; hence, power under H_3 equals power$_{33}$ = 1 − β_{23} − γ_{13}. When H_1 is true, power is reduced by an amount equal to γ_{31}; hence, power

		NATURE		
		H$_1$ true	H$_2$ true	H$_3$ true
DECISION ABOUT NATURE	Accept H$_1$	Correct $1-\beta_{21}-\gamma_{31}$ Power$_{11}$	Type 1 error α_{12}	Type 3 error γ_{13}
	Accept H$_2$	Type 2 error β_{21}	Correct $1-\alpha_{12}-\alpha_{32}$	Type 2 error β_{23}
	Accept H$_3$	Type 3 error γ_{31}	Type 1 error α_{32}	Correct $1-\beta_{23}-\gamma_{13}$ Power$_{33}$

Figure 3. Ninefold table showing three decisions (their correctness and conditional probability) for each of three states of nature. This table uses the revised definition of power.

under H_1 is power$_{11}$ = $1 - \beta_{21} - \gamma_{31}$. So the three-choice test is less powerful than the nondirectional test by an amount equal to the three-choice test's risk of Type III error, γ.

These conclusions are illustrated in Figure 4, which compares power and error rates of the three-choice test (Figure 4C) to conventional one- and two-tailed tests. Each test is represented by two sampling distributions of a sample statistic (e.g., the difference between sample means), one distribution expected under the null hypothesis (''Null'' in Figure 4) and one distribution expected under a true state of nature when the null hypothesis is false (''True'' in Figure 4). Critical values for the null distributions appear as *cv*. The shaded area of each true distribution represents power. The nondirectional test (Figure 4B) has two power areas, the large right tail and the small left tail, both representing the power decision of accepting a true H_A. In the three-choice test, only the right tail, equal in size to the right tail of the nondirectional test,

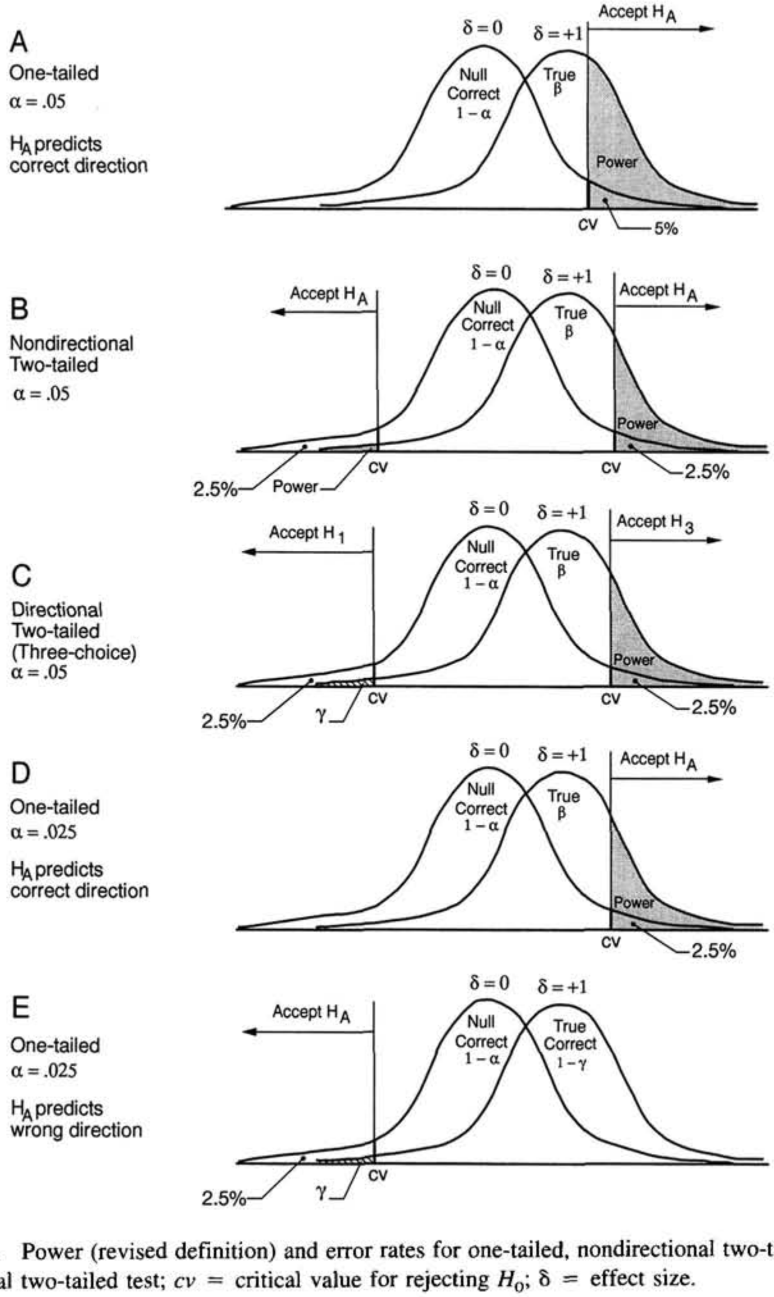

Figure 4. Power (revised definition) and error rates for one-tailed, nondirectional two-tailed, and directional two-tailed test; *cv* = critical value for rejecting H_0; δ = effect size.

represents power. The striped area in the left tail represents the probability of Type III error γ. Since γ in the three-choice test is equal in size to the left power tail of the nondirectional test, the three-choice test has less power than the nondirectional test by an amount equal to γ. Furthermore, as the true value of the parameter in the three-choice test decreases to the null value, causing the true distribution to move left, γ grows to a limit of 0.5α. Since γ is always less than 0.5α, the difference in power between the tests is always less than 0.5α (see also Kaiser, 1960, p. 164). One can put this differently by focusing on the decision rather than the test: The power of a two-tailed test will decrease by an amount less than 0.5α if one reaches a directional decision after null rejection instead of a nondirectional decision.

Three-Choice Test Versus One-Tailed Test

With the revised definition, the one-tailed test exceeds the three-choice test in power. Indeed, a one-tailed test predicting the correct direction need use only 0.5α to have the same power as a three-choice test using α (assuming α is split equally between both tails of the null distribution, i.e., $\alpha_{12} = \alpha_{32}$). For example, a one-tailed test predicting the correct direction at the 0.025 level has the same power as a three-choice test at the 0.05 level; compare Figures 4C and 4D. One choosing a one-tailed test to obtain this power advantage must make a correct directional prediction before inspecting the data and must forfeit the right, whatever the results, to decide with this test that the true direction falls opposite to the prediction. (One who inspects data in order to predict direction for a one-tailed test at α—or who conducts a one-tailed test α, finds nonsignificance, and then conducts a second one-tailed test at α predicting the opposite direction—will have conducted the equivalent of a three-choice test at 2α).

Sample Size

With the traditional definition of power, the three-choice test and the nondirectional test have the same power and therefore require the same sample size. The one-tailed test predicting the correct direction is the most powerful of the three tests and requires the smallest sample.

With the revised definition, the one-tailed test is the most powerful, requiring the smallest sample, and the three-choice test is the least powerful, requiring the largest sample; compare Figures 4A, 4B, and 4C. Figures 4B and 4C show why the nondirectional test requires a slightly smaller sample than does the three-choice test when the tests are equated on power and α. Assume for both tests that the true distribution's right power tail is exactly 0.25, which is approximately what is shown. Power for the nondirectional test is actually greater than 0.25 since power for this test is the sum of the left and right power tails. To equate the tests on power, the nondirectional test's right power tail must be reduced so that the two tails sum to 0.25. One can do this by reducing the nondirectional test's sample size. This makes the null and true distributions fatter, thereby moving the critical values farther from the null distribution's center and reducing the right power tail. As a result, for a given power and α, the nondirectional test requires a smaller sample than does the three-choice test.

While one might select the nondirectional test over the three-choice test because of the slight advantage in power and sample size, one should bear in mind that with the revised definition of power, the nature of the decision reached after null rejection—directional or nondirectional—affects the power of a two-tailed test. One planning a directional decision who selects a nondirectional test and calculates power for a nondirectional decision will inflate power over the actual value and therefore underestimate sample size. With the three-choice test, however, one calculates the actual power and sample size that result from a directional decision.

Comparing Calculations for the Directional Two-Tailed Test to Conventional Tests

With the revised definition of power, the following examples compare power, error rate, and sample size calculations for the one-tailed, nondirectional two-tailed, and directional two-tailed tests. The examples use the z test because of its mathematical simplicity. However, for the z test or for tests based on other distributions such as t or the binomial, we show how in principle one can determine power, γ, and sample size for the directional two-tailed test without calculation from statistical tables for conventional one- and two-tailed tests.

A z test for independent samples compares two sample means. The true effect (δ, where $\delta = \mu_1 - \mu_2$) is predicted to be negative, that is, μ_2 is predicted to be larger than μ_1. Assume that the smallest useful difference is 1 point. Sample sizes are $n_1 = 32$ and $n_2 = 30$; sample means are $M_1 = 40$ and $M_2 = 30$.

The population distributions are normal and homogeneous, with standard deviations $\sigma_1 = \sigma_2 = \sigma = 8$. The standard error of the difference between means ($\sigma_{M_1-M_2}$) is $\sigma_{M_1-M_2} = \sqrt{\sigma^2(1/n_1 + 1/n_2)} = \sqrt{8^2(1/32 + 1/30)} = 2.03306$. Using $\alpha = 0.05$, we calculate power and error rates for true effect $\delta = 1$, which is contrary to prediction, and calculate sample sizes needed for widely different power (π) values: 0.10, 0.30, and 0.60.

Results

Table 1 compares results for the three tests. The one-tailed test, which predicted the wrong direction because the investigator guessed wrong, was not significant and does not allow a directional decision in the correct direction. The directional and nondirectional tests, however, were significant. Unfortunately, the nondirectional test does not provide for a directional decision. Only the directional two-tailed test resulted in the correct decision that μ_1 is larger than μ_2.

Had the one-tailed test predicted the correct direction, it would have been significant and the most powerful of the tests. On the other hand, γ was larger for the one-tailed test than for the directional test. Indeed, when α is held constant, the upper limit for γ in a one-tailed test is twice that in a directional test. The reason is that the upper limit for γ is 0.5α in the directional test but is α in the one-tailed test. (For either test, γ approaches its maximum as δ approaches zero. Even though Figures 4C and 4E use different αs, note in each figure how γ approaches its maximum as the true distribution moves left toward smaller δ values.)

With the values used in the sample size computations ($\alpha = 0.05$, $\sigma = 8$, and $\delta = 1$), sample size differences between the directional and nondirectional two-tailed tests are negligible for powers of 0.60 and 0.30 and are small for a power of 0.10. Sample size differences would have been even smaller for larger effect sizes. With the values used in the power and error rate computations ($\alpha = 0.05$, $\sigma = 8$, $\delta = 1$, $n_1 = 32$, and $n_2 = 30$), power and error rate differences for the two tests are negligible and would have been even smaller for larger effect sizes. Thus, for the values used, the cost in power, error rates, or sample size of replacing the nondirectional test with the directional test would have been small to negligible.

Table 1
Summary Table Comparing Three Significance Tests Conducted on Data in Text

	One-tailed	Nondirectional two-tailed	Directional two-tailed
Hypotheses	H_0: $\mu_1 - \mu_2 \geqslant 0$ H_A: $\mu_1 - \mu_2 < 0$	H_0: $\mu_1 - \mu_2 = 0$ H_A: $\mu_1 - \mu_2 \neq 0$	H_1: $\mu_1 - \mu_2 < 0$ H_2: $\mu_1 - \mu_2 = 0$ H_3: $\mu_1 - \mu_2 > 0$
α	.05	.05	.05
Test statistic & value	$z = 4.91869$	$z = 4.91869$	$z = 4.91869$
Critical value(s)	$z_\alpha = -1.645$	$z_{\alpha/2} = \pm 1.96$	$z_{\alpha/2} = \pm 1.96$
Decision	Retain H_0	Accept H_A	Accept H_3
Power & error rates			
$\delta = 1$, $n_1 = 32$,	π = undefined[a]	$\pi = 0.078$	$\pi = 0.071$
$n_2 = 30$	β = undefined[a]	$\beta = 0.922$	$\beta = 0.922$
	$\gamma = 0.0163$[b]	γ = undefined	$\gamma = 0.007$
Sample size			
for $\pi = .10$, $\delta = 1$	n = undefined[a]	$n = 54.794 \approx 55$	$n = 58.909 \approx 59$
for $\pi = .30$, $\delta = 1$	n = undefined[a]	$n = 263.426 \approx 264$	$n = 263.788 \approx 264$
for $\pi = .60$, $\delta = 1$	n = undefined[a]	$n = 627.017 \approx 628$	$n = 627.039 \approx 628$

Note. z_α = critical value of z for one-tailed z test; $z_{\alpha/2}$ = critical value of z for two-tailed z test, both tests at α; $z = (M_1 - M_2)/\sqrt{\sigma^2(1/n_1 + 1/n_2)}$ = test statistic; n = number of subjects per group when $n_1 = n_2$ and n = harmonic mean of n_1 and n_2 when $n_1 \neq n_2$.

[a] For this one-tailed test, π, β, and n are undefined because H_A predicted the wrong direction. Had H_A predicted the correct direction, π would have been 0.124, β would have been 0.876, and n would have been 16.913, 160.816, and 461.253, respectively, for π of 0.10, 0.30, and 0.60. These calculations use standard methods (e.g., Hinkle, Wiersma, & Jurs, 1994).

[b] A Type III error is possible because H_A predicted the wrong direction.

Calculating Power and Error Rates

Power and error rate calculations in Table 1 for the directional two-tailed test and the one-tailed test are based on standard calculations for the nondirectional two-tailed test, which we discuss first.

Nondirectional Two-Tailed Test

In Figure 4B, the true distribution contains a large power area on the right and a small power area on the left. If Figure 4B conformed to the present data, standard calculational procedures (e.g., Hinkle, Wiersma, & Jurs, 1994, chap. 12) would show that the larger area is 0.071 and the smaller area is 0.007. Hence, $\pi = 0.071 + 0.007 = 0.078$. Moreover, $\beta = 1 - \pi = 1 - 0.078 = 0.922$. Note that the values of π and β calculated here for effect size $\delta = 1$ would hold also for $\delta = -1$. For the nondirectional test, it is the size, not the direction, of an effect that affects π and β.

Directional Two-Tailed-Test

Refer to Figure 4C. Type III error γ for the directional test is the true distribution's striped area on the left side, which equals in size the smaller power area of 0.007 in Figure 4B. Hence, $\gamma = 0.007$. Power π of the directional test is the true distribution's shaded area on the right side, which equals in size the larger power area of 0.071 in Figure 4B. Hence, $\pi = 0.071$. (Therefore, the difference in power between the directional and nondirectional test is equal to $\gamma = 0.007$.) Type II error $\beta = 1 - \pi - \gamma = 1 - 0.071 - 0.007 = 0.922$. The values of γ, π, and β calculated here for $\delta = 1$ would hold also for $\delta = -1$. For the directional test, it is the size, not the direction, of the effect that affects γ, π, and β. Note that since the directional two-tailed test at α and the one-tailed test at 0.5α have the same power (compare Figures 4C and 4D), standard power tables for the one-tailed test can be used to find the power of the directional two-tailed test. Here, one determines the power of a directional test at α simply by finding the power in the table for a one-tailed test at 0.5α. Moreover, γ for the directional two-tailed test at α (Figure 4C) equals the difference in power between the nondirectional two-tailed test at α (Figure 4B) and the one-tailed test at 0.5α (Figure 4D). Therefore, one can determine γ for the directional test with statistical tables for conventional one- and two-tailed tests. But first, one must verify in the table for the nondirectional test that the listed power values include the smaller power area. (The table may be providing an approximate power that does not include the smaller area.)

One-Tailed Test

If a one-tailed test predicts the correct direction, then one can compute power and β using standard methods (e.g., Hinkle et al., 1994) but γ has no meaning (see Figure 4D). If a one-tailed test predicts the wrong direction, then power and β have no meaning (because H_0 is true), but one can compute γ (see Figure 4E). For a one-tailed test predicting a given direction, both the size and the direction of an effect affects π, β, and γ. Since the present test predicts the wrong direction, we calculate only γ. Figure 4E illustrates γ for a one-tailed test using $\alpha = 0.025$. Had the present test used $\alpha = 0.025$, its γ would equal the γ of the directional two-tailed test using $\alpha = 0.05$ (compare Figures 1C and 1E). Previously calculated, that value is 0.007. Since the present test uses $\alpha = 0.05$, we calculate γ. Let z_α equal the critical value of z on the left of the null distribution ($z_\alpha = -1.645$), and let R_α equal the raw score equivalent of z_α.

Step 1. Referring to the null distribution, convert z_α to R_α: where δ_0 is δ's value under H_0 and is the null distribution's mean, $R_\alpha = \delta_0 + z_\alpha(\sigma_{M_1-M_2}) = 0 + (-1.645)(2.03306) = -3.34438$. Convert R_α to a z value under the true distribution: Where δ is the true effect size and is the true distribution's mean, $z = (R_\alpha - \delta)/\sigma_{M_1-M_2} = (-3.34438 - 1)/2.03306 = -2.1369$.

Step 2. The normal curve table shows that the area to the left of z of -2.1369 is 0.0163. Hence, $\gamma = 0.0163$.

Calculating Sample Size for a Given Power

Sample size calculations in Table 1 for the directional two-tailed test are based on standard calculations for nondirectional tests, which we discuss first.

Nondirectional Two-Tailed Test

How many subjects are needed for a power of 0.10? (We selected a small power value for the calculational example to increase the sample size differences illustrated across the three tests.) Referring to Figure 1B, the true distribution contains large and small power areas. One must find the sample size that makes these power areas sum to 10% of the distribution. Many textbooks use an approximation that ignores the smaller power area and finds the sample size that makes the larger power area by itself constitute 10% of the distribution (e.g., Guilford & Fruchter, 1978; Hildebrand, 1986; Hinkle et al., 1994; Kachigan, 1986; Kirk, 1984).

Approximation. Let n = number of subjects per group, $z_{1-\alpha/2}$ = critical value of z on the right side of the null distribution, and z_β = the z value that divides the upper 0.10 and lower 0.90 of the true distribution. From the normal distribution table, $z_{1-\alpha/2} = 1.9599$ and $z_\beta = 1.2815$. One obtains n from $n = 2[\sigma(z_\beta - z_{1-\alpha/2})/\delta]^2 = 2[8(1.2815 - 1.9599)/1]^2 = 58.909 \approx 59$. Hence, 58.909 subjects per group are required for the larger power area to constitute 10% of the true distribution. If the larger power area by itself is 0.10, then the power of the test must be greater than 0.10. Since the smaller power area must be less than 0.5α (less than 0.025), the power of the test must be less than 0.10 + 0.025 = 0.125. Hence, 58.909 subjects per group are required for power between 0.10 and 0.125.

Greater precision. Although the above approximation is adequate for many purposes, greater precision in calculating n is necessary to compare n requirements for the nondirectional and directional two-tailed tests. Greater precision can be obtained in four steps.

Step 1. Set the larger power area equal to the desired power (i.e., 0.10) and compute the smaller power area. (a) The larger power area of 0.10 results from $n = 58.909$ (see *Approximation* section), and this n produces standard error $\sigma_{M_1-M_2} = \sqrt{\sigma^2(1/n_1 + 1/n_2)} = \sqrt{8^2(1/58.909 + 1/58.909)} = 1.47406$. The $\sigma_{M_1-M_2}$ of 1.47406 produces for critical value $z_{\alpha/2}$ an equivalent raw score ($R_{\alpha/2}$) where $R_{\alpha/2} = \delta_0 + z_{\alpha/2}(\sigma_{M_1-M_2}) = 0 + (-1.95996)(1.47406) = -2.8891$. (b) The $R_{\alpha/2}$ of -2.8891 corresponds to the true distribution's z value (z_γ), where $z_\gamma = (R_{\alpha/2} - \delta)\sigma_{M_1-M_2} = (-2.8891 - 1)/1.47406 = -2.63836$, resulting in a smaller power area of 0.00417. (Thus, n of 58.909 produces a power of 0.10 + 0.00417 = 0.10417.)

Step 2. Decrease the larger power area by an amount equal to the smaller power area. Decreased larger power area is 0.10 − 0.00417 = 0.09583.

Step 3. Compute n *needed for the decreased larger power area found in Step 2.* The decreased larger power area of 0.09583 in the true distribution begins at $z_\beta = 1.30568$. Hence, $n = 2[\sigma(z_\beta - z_{1-\alpha/2})/\delta]^2 = 2[8(1.30568 - 1.95996)/1]^2 = 54.794 \approx 55$.

Step 4. Compute power produced by n *found in Step 3.* (a) The n in Step 3 produces a new standard error ($\sigma_{M_1-M_2}*$) where $\sigma_{M_1-M_2}* = \sqrt{\sigma^2(1/n_1 + 1/n_2)} = \sqrt{8^2(1/54.794 + 1/54.794)} = 1.52840$. The new standard error $\sigma_{M_1-M_2}*$ produces a new raw score critical value ($R_{\alpha/2}*$), where $R_{\alpha/2}* = \delta_0 + z_{\alpha/2}(\sigma_{M_1-M_2}*) = 0 + (-1.95996)(1.52840) = -2.99560$. (b) Compute the smaller power area resulting from the decreased larger power area found in Step 2: The $R_{\alpha/2}*$ of -2.99560 corresponds to the true distribution's z_γ, where $z_\gamma = (R_{\alpha/2}* - \delta)/\sigma_{M_1-M_2}* = (-2.99560 - 1)/1.52840 = -2.61424$, resulting in a smaller power area of 0.00447. (c) Calculate power: power = smaller area + larger area = 0.00447 + 0.09583 = 0.10030. (Thus, n of 54.794 produces a power of 0.10030.)

Repeat Steps 2–4 until useful reductions in n stop occurring in Step 3, and disregard n in Step 3 if power computed in Step 4 is less than the desired value. (Note that the size of the smaller and larger power areas, needed for Step 2 of a new iteration, is given in Step 4C of the previous iteration). Thus, precise calculations for the nondirectional test reduce n from 59 to 55.

Directional Two-Tailed Test

Refer to Figure 4C. The true distribution in the directional test has only one power area, and the directional test has a power of exactly 0.10 when this power area constitutes exactly 10% of the distribution. The n needed for this, previously calculated in the *Approximation* section for the nondirectional test, is 58.909 ≈ 59. Thus, when calculated exactly, the n required for a power of 0.10 was 59 for the directional test and 55 for the nondirectional test. Textbooks ignoring the smaller power area when approximating power of nondirectional tests would have concluded that both tests require an n of 59, thereby negating the principal advantage of nondirectional over directional two-tailed tests, namely, greater power and smaller sample size. It is important to note that the relatively simple calculations in the *Approximation* section that overestimate n for the nondirectional test produce a correct n for the directional test. Hence, an advantage of the directional over the nondirectional two-tailed test is that calculation of correct n is much easier. Indeed, since the directional two-tailed test at α and the one-tailed test at 0.5α have the same power, then other things being equal, they require the same n (compare Figures 4C and 4D); hence, standard sample size tables for the one-tailed test can be used to find a correct n for the directional two-tailed test. (An n found for the nondirectional test by this method would be approximate.) Finally, γ for the directional test equals the smaller power area of the nondirectional test when the n of 58.909 produces a larger power area of 0.10. This smaller area—γ for the directional test—was calculated in Step 1 to be .00417.

One-Tailed Test

Since H_A in the one-tailed test in Table 1 predicted the wrong direction, it is meaningless to ask what sample size is needed for a given power. No sample size will make possible the power decision of rejecting a false H_0 in favor of a true H_A. One assuming the prediction is correct can calculate sample size with standard methods (e.g., Hinkle et al., 1994, chap. 12, but see Kupper & Hafner, 1989).

Discussion

One who imposes a directional decision on a nondirectional two-tailed test will overestimate power, underestimate sample size, and make no provision for assessing Type III error if traditional calculations—those that apply to nondirectional decisions—are used. These errors can be small to trivial for the z test, as Table 4 illustrates. But for certain conditions—for example, when α is large and effect size is small—the errors might be important. Moreover, the z test uses the normal distribution, which is a symmetrical distribution having thin tails. With significance tests based on other distributions, the errors may be larger than those for the z test.

Whatever their size, it makes little sense to tolerate the errors when a directional two-tailed test could have been conducted from the outset. The change to the directional test requires only restructuring the original statistical hypotheses and calculating power, error rates, and sample size for directional rather than nondirectional decisions. Indeed, a case can be made that one imposing a directional decision on a nondirectional test is actually conducting a directional two-tailed test, but conducting it poorly: The statistical hypotheses were incorrectly structured as nondirectional rather than directional, and power, error rates, and sample size were incorrectly calculated for nondirectional rather than directional decisions.

In comparison with the nondirectional test, the directional test (a) provides for reaching a directional decision, (b) provides power and sample size calculations for directional decisions, (c) provides for determining the risk of getting the direction wrong, (d) offers much easier precise calculation of sample size, (e) is conceptually simpler when a directional decision is planned, and (f) will be less confusing to students. The nondirectional test confuses students because if a directional decision is reached, they cannot understand the statistical reasoning by which the testing of nondirectional hypotheses has led to a directional decision and, if a nondirectional decision is reached, they cannot understand why the nondirectional test is used at all since most interesting scientific questions about effects or phenomena require information about direction for a satisfactory answer. In comparison with the nondirectional test, the disadvantage of the directional test in practical situations is a slight to trivial decrease in power (always less than 0.5α) and a corresponding increase in sample size. But the sample size advantage of the nondirectional z test will not materialize if one calculates n for a nondirectional decision with the approximation procedure used by many textbooks.

Confidence Intervals

One can use a confidence interval to assess direction. Direction can mean either the sign of a parameter (positive or negative) or the direction of a parameter's discrepancy from an interesting nonzero value (larger or smaller than the value). But many researchers simply will not use a confidence interval. In this event, we recommend the directional test over the other significance tests. For researchers willing to consider a confidence interval, the decision between the confidence interval and the directional test is more complex than simply choosing one to the exclusion of the other. We will argue for adopting a confidence interval as the primary technique for evaluating direction—with a directional test, or parts of it, computed as an adjunct.

We begin by discussing how to investigate parameter sign. Using the above data, one can determine the sign of the $\mu_1 - \mu_2$ difference with a two-sided 95% confidence interval (95% CI), for which the limits are found from 95% CI $= (M_1 - M_2) \pm z_{0.025}(\sigma_{M_1-M_2}) = (40 - 30) \pm 1.959(2.033) = 10 \pm 3.983$. The interval extends from 6.017 to 13.983, and since it contains only positive values, one can be at least 95% confident that $\mu_1 - \mu_2$ is positive, that is, that μ_1 is larger than μ_2. (If the interval had contained zero, the results would have been inconclusive regarding direction.) Since $\delta = 1$ (see above), the interval results in a correct decision about the sign. One can also investigate the direction of a parameter's discrepancy, for example, when investigating where μ is larger or smaller than a predicted value of 5. If all interval values are smaller than 5, one decides with confidence at least equal to the confidence coefficient that μ is smaller than 5. If all interval values are larger than 5, one decides with the same confidence that μ is larger than 5. If the interval contains 5, one decides that the results are inconclusive regarding the direction of any discrepancy. (The directional test addressing this problem would evaluate H_1: $\mu < 5$, H_2: $\mu = 5$, and H_3: $\mu > 5$.) To save space, the following discussion on one-sided intervals deals only with parameter sign. But it also applies, with minor changes in wording, to parameter discrepancy.

One can assess direction with one-sided confidence intervals. There are two kinds of one-sided intervals, one having a lower limit but no upper limit and one having an upper limit but no lower limit (e.g., Kirk, 1984, p. 322). In either case, if all values in the interval contain the same sign, one concludes with confidence at least equal to the confidence coefficient that the sign in the interval indicates the true sign. One decides which one-sided interval to use by predicting the direction of the sign before inspecting the data, similar to selecting the direction for the alternative hypothesis in a one-tailed significance test. (Inspecting the data to predict direction will lower the confidence coefficient from its preselected value.) When a negative sign is predicted, one uses an upper limit interval. Only the upper limit interval can contain uniformly negative values and result in the decision that the sign is negative. Unfortunately, an upper limit interval can never contain uniformly positive values; hence, when the prediction is incorrect and the true sign is positive, the correct decision that the sign is positive can never result. Similarly, when a positive sign is predicted, one uses a lower limit interval. But when the prediction is incorrect, the correct decision that the sign is negative can never result. Thus, one-sided intervals "predicting" the wrong direction cannot detect the true direction. For example, although $\delta = 1$ for the above data, a negative sign was predicted. So we calculate an upper limit interval. The 95% upper limit confidence interval (95% $\mathrm{CI}_{\mathrm{UL}}$), for which the upper limit is found from 95% $\mathrm{CI}_{\mathrm{UL}} = (M_1 - M_2) + z_{0.05}(\sigma_{M_1-M_2}) = (40 - 30) + 1.645(2.033) = 10 + 3.344 = 13.344$, produces an interval that contains the upper limit 13.344 and all smaller values (including all negative values). Since the interval contains zero, the results are inconclusive regarding direction. In comparison with two-sided intervals, one-sided intervals have a higher probability of Type III error when predicting the wrong direction but a higher probability of identifying the true direction when predicting the correct direction.[2] Arguments regarding the choice of one- versus two-sided confidence intervals for detecting direction closely parallel arguments in the historical debate regarding one- versus two-tailed significance tests (e.g., Braver, 1975; Burke, 1953, 1954; Cohen, 1965; Eysenck, 1960; Goldfried, 1959; Gravetter & Wallnau, 1992, p. 228; Hick, 1952; Jones, 1952, 1954; Kimmel, 1957; Marks, 1951, 1953; Pillemer, 1991; Welkowitz, Ewen, & Cohen, 1991, pp. 150–151). We recommend two-sided intervals for most problems in which identifying direction is the main objective, mainly because one-sided intervals can detect only a correctly predicted direction.

Should one use a two-sided confidence interval or a directional two-tailed test to make a directional decision? It makes no difference: a directional two-tailed test at α and a two-sided confidence interval using confidence coefficient $1 - \alpha$ will result in the same decision about direction. (However, there are problems when the parameter is a population proportion; see Koopmans, 1987, pp. 281–282.) But more than a decision about direction may be of interest to the investigator or to the scientific community, and a confidence interval provides information not provided by a significance test (e.g., Natrella, 1960): The $1 - \alpha$ two-sided confidence interval (a) consists of all values for the null hypothesis that would lead to null retention in a (directional or nondirectional) two-tailed test at α, (b) estimates the direction of the parameter with at least $1 - \alpha$ confidence, (c) estimates the signed size (as opposed to absolute size) of the parameter with exactly $1 - \alpha$ confidence, and (d) indicates the precision of the estimate with the width of the interval. Furthermore, since a confidence interval provides an estimate of the parameter given the obtained data, it avoids a decades-old criticism of significance tests that argues that, although researchers seek the conditional probability that a hypothesis about the parameter is true given the obtained data, significance tests provide instead the conditional probability of the obtained data given that a (point) hypothesis about the parameter is true (e.g., Bakan, 1966; Carver, 1978; Cohen, 1994). We agree with this criticism but point out that the information yield of a significance test consists of attaching a conditional probability not only to the obtained data given a point hypothesis about the parameter but also to the cells in the fourfold and ninefold tables. The conditional probability for a cell is the probability that a certain decision about a statistical hypothesis will be reached given some true state of nature expressed with a point value. Note, however, that although a researcher may reach a "retain," "accept," or "reject" decision about a statistical hypothesis, nothing in a significance test—not

[2] A Type III error occurs in a confidence interval when all values in the interval contain the same sign, leading to a directional decision, but the true sign has the opposite direction.

even the *p* value—provides the probability or confidence that the decision is correct or that the statistical hypothesis is true. (A two-tailed *p* value is the conditional probability of obtaining data as extreme or more extreme in either direction as the observed data, given that the null hypothesis is true.) In contrast, because a confidence interval's estimate of the parameter is conditioned only on the obtained data, the confidence coefficient for the interval comes as close as we are likely to get to an unconditional probability or confidence that a decision about the parameter's direction or size is correct. Thus, a two-sided confidence interval provides information not provided by a directional two-tailed test, namely, an estimate of parameter direction and size given the data, the precision of the estimate, and all the null values that would lead to null retention in the test.[3] Accordingly, we recommend the two-sided confidence interval to evaluate the direction or size of a parameter.[4]

However, one employing a confidence interval may find significance test information to be useful as well. For example, since a directional two-tailed test at α and a $1 - \alpha$ two-sided confidence interval suffer the same risk of Type III error, one reaching a directional decision with a confidence interval can determine its Type III error risk with the above methods for the directional test. Another example: Suppose one computes a confidence interval to determine whether it contains a theoretically predicted point value. One might also compute a directional test's two-tailed *p* as a measure of the consistency between the theoretical prediction, expressed by the test's null hypothesis, and the obtained data. (See Thompson, 1987, who also endorses supplementing confidence intervals with *p* values and Gibbons & Pratt, 1975, for problems with two-tailed *p* values when the null distribution is asymmetrical or discrete.) Thus, a directional test provides information not provided by a two-sided confidence interval. So one computing a confidence interval might consider also computing a directional test, or parts of it, as an adjunct (but not as a determiner of whether to compute the interval—see Footnote 3).

Conclusion

In summary, our endorsement of the directional two-tailed test over conventional significance tests should not be construed as an endorsement of it in particular, or of significance tests in general, as the principal methodological tool of psychology. Strong arguments have been brought against significance tests (e.g., Carver, 1978; Cohen, 1994; Dar, 1987; Meehl, 1967, 1978; Morrison & Henkel, 1970; Schmidt, 1992), and we agree with many of them. So where it is possible to decide direction with a confidence interval or significance test, we recommend considering the two-sided confidence interval for the primary analysis. But we also recommend the directional two-tailed test, or parts of it, as an adjunct. The directional test provides additional information the investigator might find useful. If one decides to use a significance test alone, we recommend the directional two-tailed test as the best choice for most purposes. If one uses the nondirectional test but plans a directional decision, we suggest using the procedures outlined above to determine power, error rates, and sample size.

[3] Researchers sometimes attempt to cover all bases by conducting a significance test and, when results are significant, calculating a confidence interval to estimate parameter size. Unfortunately, Schmidt (1992) has shown that making the decision to estimate parameter size contingent on the outcome of the significance test produces a biased estimate of the parameter. Specifically, over many replications of the same study, this procedure produces estimates whose average value exceeds the true parameter size, with the error increasing as true size decreases. When true size decreases below the critical value needed for null rejection, a single study will necessarily overestimate the parameter since only data equal to or larger than the critical value (hence, larger than the parameter) will produce significance and be used for the estimate.

[4] We leave the door open to changing our recommendation of confidence intervals to decide direction. Since the $1 - \alpha$ two-sided confidence interval and the directional test at alpha make the same directional decision, then (a) if the directional test rejects the null in favor of a particular direction, the confidence interval will contain values uniformly having that direction and (b) if the test does not reject the null, the interval will not contain values having a uniform direction. Hence, because the directional test can predict the directional decision made by a confidence interval, the test provides the same ''unconditional'' confidence provided by the confidence interval that the directional decision is correct. Thus, the directional test becomes roughly equal to the confidence interval for deciding direction. This argument will be controversial and we need to study it further before formally proposing it.

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Braver, S. L. (1975). On splitting the tails unequally: A new perspective on one- versus two-tailed tests. *Educational and Psychological Measurement, 35,* 283–301.

Burke, C. J. (1953). A brief note on one-tailed tests. *Psychological Bulletin, 50,* 384–387.

Burke, C. J. (1954). Further remarks on one-tailed tests. *Psychological Bulletin, 51,* 587–590.

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

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

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

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist, 42,* 145–151.

Eysenck, H. J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review, 67,* 269–271.

Ferguson, G. A., & Takane, Y. (1989). *Statistical analysis in psychology and education* (6th ed.). New York: McGraw-Hill.

Gibbons, J. D., & Pratt, J. W. (1975). *P*-values: Interpretation and methodology. *The American Statistician, 29,* 20–25.

Goldfried, M. R. (1959). One-tailed tests and 'unexpected' results. *Psychological Review, 66,* 79–80.

Gravetter, F. J., & Wallnau, L. B. (1992). *Statistics for the behavioral sciences* (3rd ed.). New York: West.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York: McGraw-Hill.

Hand, J., McCarter, R. E., & Hand, M. R. (1985). The procedures and justification of a two-tailed directional test of significance. *Psychological Report, 56,* 495–498.

Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart & Winston.

Hick, W. E. (1952). A note on one-tailed and two-tailed tests. *Psychological Review, 59,* 316–318.

Hildebrand, D. K. (1986). *Statistical thinking for behavioral scientists.* Boston: Duxbury Press.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1994). *Applied statistics for the behavioral sciences* (3rd ed.). Boston: Houghton Mifflin.

Hodges, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, London, Series B (Methodological), 16,* 261–268.

Hopkins, B. (1973). Educational research and Type III errors. *The Journal of Experimental Education, 41,* 31–32.

Jones, L. V. (1952). Tests of hypotheses: One-sided vs. two-sided alternatives. *Psychological Bulletin, 49,* 43–46.

Jones, L. V. (1954). A rejoinder on one-tailed tests. *Psychological Bulletin, 51,* 585–586.

Kachigan, S. K. (1986). *Statistical analysis: An interdisciplinary introduction to univariate & multivariate methods.* New York: Radius Press.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review, 67,* 160–167.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kiess, H. O. (1996). *Statistical concepts for the behavioral sciences* (2nd ed.). Boston: Allyn and Bacon.

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin, 54,* 351–353.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Kirk, R. E. (1984). *Elementary statistics* (2nd ed.). Monterey, CA: Brooks/Cole.

Koopmans, L. H. (1987). *Introduction to contemporary statistical methods* (2nd ed.). Boston: Duxbury Press.

Kupper, L. L., & Hafner, K. B. (1989). How appropriate are popular sample size formulas? *The American Statistician, 43,* 101–105.

Marks, M. R. (1951). Two kinds of experiment distinguished in terms of statistical operations. *Psychological Review, 58,* 179–184.

Marks, M. R. (1953). One- and two-tailed tests. *Psychological Review, 60,* 207–208.

May, R. B., Masson, M. E. J., & Hunter, M. A. (1990). *Application of statistics in behavioral research.* New York: Harper & Row.

McClave, J. T., & Sincich, T. (1995). *A first course in statistics* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

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,* 806–834.

Minium, E. W., King, B. M., & Bear, G. (1993). *Statistical reasoning in psychology and education* (3rd ed.). New York: Wiley.

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

Mosteller, F. (1948). A k-sample slippage test for an extreme population. *Annals of Mathematical Statistics, 19,* 58–65.

Natrella, M. G. (1960). The relation between confidence intervals and tests of significance. *American Statistician, 14,* 20–22, 33.

Nosanchuk, T. A. (1978). Serendipity tails: A note on two tailed hypothesis tests with asymmetric regions of rejection. *Acta Sociologica, 21,* 249–253.

Pagano, R. R. (1994). *Understanding statistics in the behavioral sciences* (4th ed.). St. Paul, MN: West.

Pillemer, D. B. (1991). One- versus two-tailed hypothesis tests in contemporary educational research. *Educational Researcher, 20,* 13–17.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

Shaffer, J. P. (1972). Directional statistical hypotheses and comparisons among means. *Psychological Bulletin, 77,* 195–197.

Spatz, C., & Johnston, J. O. (1989). *Basic statistics: Tales of distributions* (4th ed.). Pacific Grove, CA: Brooks/Cole.

Thompson, W. D. (1987). Statistical criteria in the interpretation of epidemiologic data. *American Journal of Public Health, 77,* 191–194.

Welkowitz, J., Ewen, R., & Cohen, J. (1991). *Introductory statistics for the behavioral sciences* (4th ed.). New York: Harcourt Brace Jovanovich.

Received April 4, 1995
Revision received December 17, 1995
Accepted March 14, 1996 ■