Order Number 9423728

The comparative power of several nonparametric alternatives to the Analysis of Variance test for interaction in a 2 x 2 x 2 layout

Kelley, Deborah Lynn, Ph.D.

Wayne State University, 1994

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

THE COMPARATIVE POWER OF SEVERAL NONPARAMETRIC ALTERNATIVES TO THE ANALYSIS OF VARIANCE TEST FOR INTERACTION IN A 2 x 2 x 2 LAYOUT

by

DEBORAH LYNN KELLEY

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

1994

MAJOR: EVALUATION AND RESEARCH

Approved by:

Shlomo Sawilowsky 2/16/94

Advisor Date

Dee Carruth

Donald R. Marcotte

Jim L. Pounder

R. Craig Roney

DEDICATION

This dissertation is dedicated to the memory of my hero--my mother.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

OVERVIEW

Introduction

Parametric statistics are those which make assumptions about the population from which the data were obtained (Siegel, 1956). A relatively new branch of statistics has been the development of nonparametric statistics, which was defined by Hollander and Wolfe (1973) as, "a statistical procedure that has (certain) desirable properties that hold under relatively mild assumptions regarding the underlying population(s) from which the data are obtained" (p. 1).

A detailed review of the history of both of these branches of modern statistics is outlined in Chapter Two of this study. However, briefly, nonparametric statistics gained popularity during the 1950s (Blum & Fattu, 1954; Siegel, 1956; Tate & Clelland, 1957), stabilized during the 1960s (Page & Marcotte, 1966), then experienced a decline for three reasons which were given by Anderson (1961) and Blair (1981), and summarized by Sawilowsky (1990).

> First, it is usually asserted that parametric statistics are extremely robust with respect to the assumption of population normality (Boneau, 1960; Box, 1954; Glass et al., 1972; Lindquist, 1953), precluding the need to consider alternative tests. Second, it is assumed that nonparametric tests are less powerful than their parametric counterparts (Kerlinger, 1964, 1973; Nunnally, 1975), apparently regardless of the shape of the population from which the data were sampled. Third, there has been a paucity of nonparametric tests for the more complicated research designs. (Bradley, 1968)

Each of these three reasons are discussed in detail in the next chapter. However, this study specifically addresses the third reason given above, which is the lack of nonparametric tests for complex research designs.

Two of the most common nonparametric alternatives to the parametric Analysis of Variance (ANOVA) are the Kruskal-Wallis and Friedman tests (Ingram & Monks, 1992). These tests, by themselves, are appropriate as long as an interaction effect is not present in the experiment. However, interaction effects are fairly common (Tate & Clelland, 1957). Toothaker and Chang (1980) noted, "suitable tests for main effects and interactions were needed to complete the package of those methods not making the normality assumption" (p. 169). Anderson (1961) noted that the interaction effect is the most important effect for psychologists, and stated, "few nonparametric tests of interaction exist" (p. 314).

Ten new nonparametric tests for interaction have been developed over the past 25 years, most within the last 10 years. These ten tests were summarized by Sawilowsky (1990), who advised that additional studies were necessary prior to using many of the ten techniques. Harwell (1991) underscored the need for research in this area by stating, "The interpretation of interactions using ranks has raised a number of concerns...These issues have by no means been resolved, and researchers would be well advised to exercise caution in interpreting rank interactions" (p. 391).

The Purpose of this Study

The purpose of this study is to use Monte Carlo techniques to compare the power of the Analysis of Variance (ANOVA) test for interaction to the following nonparametric statistics: (a) Bradley's Collapsed and Reduced Technique, (b) Harwell and Serlin's L test in trace criterion form, and (c) Blair and Sawilowsky's Adjusted Rank Transform test. The tests were examined in the context of a 2 x 2 x 2 layout. Data was generated from four theoretical distributions and sampled with replacement from two real data sets. Type I error rates and power properties of the four competing tests were evaluated and are presented in Chapters Four and Five of this study.

Definitions

In the stated purpose of this study, the following terms are defined.

Interaction

"Interaction is present when the pattern of differences associated with either one of the independent variables change as a function of the levels of the other independent variable" (Keppel & Zedeck, 1989, p. 187). An interaction of higher order occurs in an experimental design which has three or more effects. In a 2 x 2 x 2 factorial layout, the following effects are possible: (a) main effects (A, B and C), (b) lower order interactions (A x B, A x C, and B x C), and (c) higher order interaction (A x B x C) (Kerlinger, 1973).

Robustness

Berenson and Levine (1992) defined robust procedures as procedures which are, "relatively insensitive to slight violations in the assumptions" (p. 415). For the purposes of this study, Bradley's (1978) liberal criterion for robustness will be adopted, which is defined by Bradley as $.05\alpha \leq \pi \leq 1.5\alpha$, where α refers to the nominal significance level, and π which refers to the actual Type I error rate. Therefore, for a nominal .05 alpha level, the p value would range from .025 to .075.

Power

Power is defined as, "The probability of rejecting the null hypothesis when it is in fact false" (Runyon & Haber, 1991, p. 452). The formula for power is P = 1 - ß, where ß is equal to the actual Type II error probability (Ito, 1980, p. 209).

Asymptotic Relative Efficiency (ARE)

The ARE (also known as the Pitman efficiency) is a ratio between two comparative tests. Blair and Higgins (1985) defined the ARE as, "the limiting value of b/a as a is allowed to vary in such a way as to give test A the same power as test B while b approaches infinity and the treatment effect approaches zero" (p. 120). Generally, the efficiency of the competing nonparametric statistic is divided by that of the parametric statistic. Therefore, if the ratio is found to be less than one, the comparative nonparametric test is predicted to be less powerful than the parametric test. Conversely, if the ratio is found to be greater than one, the comparative nonparametric test is predicted to be more powerful than the parametric test.

The ARE is a rough predictor of power to be used when the sample sizes are large. It is widely used because it can compare tests under standardized conditions (Bradley, 1968).

Research Problem

This study investigated the comparative robustness with respect to departures from normality (i.e., Type I error) and power properties of four tests of interaction in the 2 x 2 x 2 ANOVA layout. The four tests are: (a) Analysis of Variance, (b) Bradley's Collapsed and Reduced Technique, (c) Harwell and Serlin's L test, and (d) Blair and Sawilowsky's Adjusted Rank Transform test.

Relevance to Education and Psychology

When conducting research, it is important that the educator and psychologist choose the most powerful test available for the specified purpose of the study. Research is frequently used to aid in the decision making process (Berenson & Levine, 1992), and therefore, decisions which affect the economic, social, political and programmatic areas of education and psychology require tools which have known power properties.

Nonparametric statistics experienced an increase in use in the 1950s, followed by a decline in use in the 1970s, and have slowly gained credibility since the 1980s. The earliest nonparametric tests for interactions were introduced 25 years ago. However, most nonparametric tests for interactions have only been developed within the last ten years. Because of the recent introduction of these nonparametric tests for interactions, several of the new tests have not been thoroughly explored, particularly in the context of a Monte Carlo study. This study investigated the robustness and power properties of selected nonparametric tests for interaction and provides information about the strength of their use. This will assist the data analyst in education, psychology and other applied related fields.

Limitations of the Study

The limitations of this study are listed below.

1. Although it is recognized that there are numerous mathematical distributions as models for real data available for investigation, this study is limited to the following: (a) Gaussian, (b) uniform, (c) $\underline{t}$ (with three degrees of freedom), and (d) exponential.
2. This study is limited to the following two distributions of real data provided by Micceri (1989): (a) multi-modal lumpy, and (b) discrete mass at zero.

3. The parametric comparison is limited to shift of location in a 2 x 2 x 2 layout.

CHAPTER TWO

THEORETICAL FOUNDATIONS AND LITERATURE REVIEW

Analysis of Variance

Over the past fifty years, Analysis of Variance (ANOVA) has become one of the foundations of statistical methods. Kerlinger (1973) stated, "modern statistical methods culminate in analysis of variance, multiple regression analysis, and factor analysis" (p. 216). ANOVA is used to determine if there is an overall difference among the group means due to the treatment (Runyon & Haber, 1991). Whereas the t test is used to compare just two groups, the use of ANOVA is vital because it is used to compare more than two groups, which is frequently necessary in education, psychology, and other related disciplines (Kerlinger, 1973).

This paper is limited to ANOVA in comparison to nonparametric alternatives, and therefore, does not examine Analysis of Covariance (ANCOVA). For viewpoints related to ANCOVA comparisons to nonparametric alternatives, see Conover and Iman (1981, 1982), Harwell and Serlin (1988), Olejnik and Algina (1984).

Analysis of Variance in an A x B x C Design

ANOVA in an A x B x C design is called a factorial design, in that "Factorial analysis of variance is the statistical method that analyzes the independent and interactive effects of two or more independent variables on a dependent variable" (Kerlinger, 1973, p. 245). In an A x B x C design, there are three independent variables, and one dependent variable. Assuming that each of the three independent variables form a 2 x 2 x 2 layout, this arrangement provides for seven hypotheses. The first three hypotheses test for main effects: (a) the differences between A1 and A2, (b) the differences between B1 and B2, and (c) the differences between C1 and C2. The last four hypotheses test for interaction effects: (a) (A x B), (b) (A x C), (c) (B x C), and (d) (A x B x C).

Analysis of Variance Test for Interaction

According to Keppel and Zedeck (1989), "Interaction is present when the pattern of differences associated with either one of the independent variables changes as a function of the levels of the other independent variable" (p. 187). In discussing a 2 x 2 design, Kerlinger (1973) presented the example of teaching methods and student motivation as independent variables, and achievement as the dependent variable. If the two methods operating by themselves do not differ in their effect and if types of student motivation do not differ in their effect, there are no main effects. Interaction occurs when the two independent variables join together to create an effect.

> When methods and types of motivation are allowed to work together, when they are permitted to interact, they are significantly effective...then we could come to the likely conclusion that there was an interaction between the two variables in their effect on the dependent variable (Kerlinger, 1973, p. 250).

This effect is defined as an interaction effect.

In the regression paradigm, ANOVA detects interaction effects by using techniques such as partialing, and ultimately forming F ratios for the effects analyzed (Keppel & Zedeck, 1989). The corresponding statistical hypotheses for a 2 x 2 factorial layout are outlined as follows by Keppel and Zedeck (1989).

> For the main effects, the null hypothesis states that the population treatment means corresponding to the separate main effects are the same; the alternative hypothesis states that they are not all equal. For the interaction, the null hypothesis states that interaction

> effects are completely absent in the population; the alternative hypothesis states that they are not.... A significant F indicates that the null hypothesis is untenable and that we should accept the alternative hypothesis that a particular factorial effect--A main effect, B main effect, or interaction--is present. (p. 196)

Analysis of Variance Assumptions

There are four major assumptions in the use of ANOVA (Cochran, 1947; Winer, 1971). Mansfield (1986) cautioned, "If these assumptions are not met, the use of the analysis of variance may be misleading" (p. 436-437). Not content with one admonition, Mansfield (1986) reiterated, "To repeat, the analysis of variance should not be used unless the relevant assumptions are at least approximately fulfilled" (p. 437).

The first assumption is independence of errors. In ANOVA, error refers to the "difference of each value from its own group mean" (Berenson & Levine, 1992, p. 506). According to Becker and Harnett (1987), independence of error means that, "no error can be predicted on the basis of any other error" (p. 401). Glass, Peckham, and Sanders (1972) advised, "Nonindependence of errors can have serious effects on the validity of probability statements in the $\underline{t}$ test or ANOVA" (p. 512).

The second assumption is that the components that comprise the total variance are additive. Many authors of statistics textbooks do not include this assumption, as it is usually self evident (Berenson & Levine, 1992; Keppel & Zedeck, 1989).

The third assumption is homogeneity of variance. "Because the 'within group,' or error calculation is based on the sum of variances within the individual groups, this calculation may be significantly impacted by unequal variances" (Berenson & Levine, 1992, p. 505).

The fourth assumption is that the scores in each group are normally distributed.

Berenson and Levine (1992) noted, "as long as the distributions are not extremely different from the normal distribution, the level of significance of the analysis of variance test is not greatly affected by lack of normality, particularly for large samples" (p. 505). However, many times the scores are not normally distributed, as Harwell (1988) stated,

> The need for analytic techniques that remove the normality requirement is apparent when researchers realize that (a) many psychological and educational data are unimodal but quite flat or skewed (i.e., non-normal) in shape, and (b) under this condition, PAR [parametric] tests often have poor distributional properties in comparison to their NPAR [nonparametric] counterparts. (p. 35)

Because of the underlying assumptions, ANOVA falls into the parametric branch of statistics. Both parametric and nonparametric statistics are discussed below.

Parametric and Nonparametric Tests Defined

Runyon and Haber (1991) defined a nonparametric test as follows.

> A nonparametric test of significance is defined as one that makes no assumptions concerning the shape of the parent distribution or population, and accordingly is commonly referred to as a distribution-free test of significance. (p. 467)

Mansfield (1986) stated, "The hallmark of these tests is that they avoid the assumption of normality" (p. 383). Parametric tests, "depend on assumptions about the parameters of the population" (Becker & Harnett, 1987, p. 576).

Berenson and Levine (1992) outlined three characteristics of parametric or classical procedures.

> First, they require that the level of measurement attained on the collected data be in the form of an interval scale or ratio scale. Second, they involve hypothesis testing of specified parameters...Third, classical procedures require very stringent assumptions and are valid only if these assumptions hold. Among

these assumptions are

1. That the sample data be randomly drawn from a population that is normally distributed.

2. That the observations be independent of each other.

3. For situations concerning central tendency for which two or more samples have been drawn, that they be drawn from normal populations having equal variances. (pp. 551-552)

It should be noted that there has been some controversy regarding the appeal to Stevens' (1946) scales of measurement, which are referred to in the quote above when discussing interval and ratio data (Gaito, 1960), as a requirement for parametric statistics. Stevens (1946) presented the following four hierarchical categories into which variables are designated: (a) nominal, (b) ordinal, (c) interval, and (d) ratio. According to Baker, Hardyck and Petrinovich (1966), Stevens' scales of measurement are the most widely accepted measurement scales, as noted, "At least two current statistics texts intended for psychologists (Senders, 1958; Siegel, 1956) present this view as gospel" (Baker, et al., 1966, p. 291). Senders (1958) devoted chapters 3-14 entirely to Stevens' scales of measurement, outlining the appropriate statistical analysis to be used for each level of measurement. Others (e.g., Anderson, 1961), have argued that the selection of the statistic should not be influenced by the scale of measurement. The issue is relevant to this study because some of the statistics in this study are based on ranking data (i.e., rank transformation). When using rank transformation, the original interval or ratio data will be transformed to the lower level ordinal data, and thus is no longer quantitative data under Stevens' scales of measurement. Harwell (1991) summarized this discussion in the following quote.

> The interpretation of interactions using ranks has raised a number of concerns (Bradley, 1968, p. 22). Among these are the relationship between the original scores and their ranks and level of measurement concerns. A key issue is whether it is meaningful to define an interaction as a difference of differences in rank means...These issues have by no means been resolved, and researchers would be well advised to exercise caution in interpreting rank interactions (p. 391).

For further information on this topic, refer to Sawilowsky (1993).

From the definition of parametric statistics given above, it is obvious that ANOVA falls into the parametric category. However, there are occasions when experimental situations and the resulting data violate parametric requirements, particularly normality of distribution. As Boneau (1960) noted, "As psychologists who perform in a research capacity are well aware, psychological data too frequently have an exasperating tendency to manifest themselves in a form which violates one or more of the assumptions underlying the usual statistical tests of significance" (p. 49). "Thus, if the populations are very far from normality, statisticians often prefer nonparametric techniques" (Mansfield, 1986, p. 395).

Because nonparametric techniques are frequently preferred when populations are not normally distributed, the question becomes, how unusual is it that populations are nonnormal? Micceri (1989) humorously referred to this issue in the title of his study, "The Unicorn, The Normal Curve, and Other Improbable Creatures." This study analyzed 440 large-sample achievement and psychometric measures, and determined all to have nonnormal distributions based on the Kolmogorov-Smirnov test of normality at the .01 alpha level. In addition, in the field of psychology, Nunnally (1978) commented, "Strictly speaking, test scores are seldom normally distributed" (p. 160).

How Rank-Based and Rank Transform Tests Work

Several of the tests for interactions discussed in this study are rank-based and may require a rank transformation; therefore a brief discussion on rank-based tests and the rank transform is necessary. In the rank transform, "the smallest observation is usually assigned the integer value 1, the next smallest 2, and so on. Any subsequent analysis is performed on the ranks" (Harwell, 1988, p. 35). In the two sample case, a t test on the ranks is equal to the Wilcoxon Rank-Sum test on the original scores. Conover and Iman (1981) presented rank transformation procedures as a "useful tool for developing nonparametric procedures to solve new problems" (p. 124). The primary issue in this subject develops when the data are nonnormal. When analyzing a contaminated normal population with both block and interaction effects, Iman found that "the rank transform approach maintains its superior power in detection of both effects" (p. 232).

To complete the literature review on this subject, it should be noted that in the context of the paired samples t test, Blair and Higgins (1985) demonstrated that the rank transform was invalid. In Monte Carlo studies conducted by Sawilowsky and Blair (1989), it was demonstrated that the rank transform is invalid for testing interactions in the 2 x 2 x 2 ANOVA layout. Thus, Sawilowsky, Blair and Higgins (1989) advised, "that researchers avoid this test except in those specific circumstances where its properties are well understood" (p. 255). Thompson (1991) reviewed the theoretical efficiency of the rank transform procedures and found the rank transform to have no theoretical base. Akritas (1991) examined the asymptotic properties of the rank transform and found, "with the exception of one testing problem, the rank transform procedure is not generally applicable" (p. 457).

Although not considered in this study, another type of nonparametric test is based on permutation methods. In some cases, the permutation method has actually proven to be slightly more powerful than parametric alternatives; however, "Permutation and randomization tests are not widely applied. In addition to the numerical burden, they are currently limited to fairly simple designs" (Welch, 1990, p. 693). However, a cursory review of the literature in the last three years shows increasing sophistication of computer hardware, making permutation use less difficult.

Historical Perspective of Parametric vs Nonparametric Tests

As outlined above, one of the assumptions of a parametric test is the assumption of normality. The normal distribution was identified by Carl Gauss (1777-1855). Gauss was a German mathematician and astronomer who noticed a recurring pattern of errors in repeated measurements. This pattern was called the Gaussian Distribution, and today is known by several other names, including the normal curve (Ingram & Monks, 1992). The normal curve was found to fit the stature of a group of Belgian soldiers, and in 1889, Galton used it extensively (Pearson & Please, 1975). In reviewing both Fisher's (1935) work, and Gosset's (Student, 1908a, 1908b) work, it is clear that both relied heavily on the assumption of normality. In discussing samples drawn from the same normally distributed population, Fisher (1935) stated, "This is the type of null hypothesis which experimenters, rightly in the author's opinion, usually consider it appropriate to test, for reasons not only of practical convenience, but because the unique properties of the normal distribution make it alone suitable for general application" (p. 45). The normal curve is still held in high regard by many, as Glass and Stanley (1970) indicated, "Indeed, the wide application and occurrence of the normal distribution are a wonder" (p. 102).

In spite of overwhelming acceptance of the normal curve by some, over the years, there have been those who doubted its widespread prevalence. As long ago as the 1940s, Geary (1947) raised doubts about the assumption of normality and the use of standard tables when normality is not assured. Geary (1947) stated,

> Standard tables cannot validly be used unless tests, based on the sample from which the inferences are to be drawn, or on a series of samples produced under similar conditions, have established the likelihood that the universal distribution is approximately normal. In certain cases--but these must be few--the nature of the material may, of itself, suffice to justify the assumption of universal normality. (p. 239)

A forbearer to nonparametric statistics came from Hotelling and Pabst (1936) who published a paper on rank correlation, and Fisher and Yates (1938) who first proposed the use of ANOVA tests on rank score data because they were dealing with original ordinal data. Beginning in the mid-1950s, nonparametric statistics gained popularity in part because of the concern of the performance of parametric procedures when faced with data from nonnormal distributions (Hollander & Wolfe, 1973). Blum and Fattu (1954) noted, "Before 1953 there were no educational statistics books that seriously concerned themselves with nonparametric statistics" (p. 469). Another reason for the increased popularity of nonparametric statistics was the emergence of new publications in the 1950s which gave simplified explanations of the workings of nonparametric statistics. These publications, along with Siegel's (1956) textbook, helped those with limited mathematical training understand and learn to use nonparametric statistics (Blum & Fattu, 1954).

According to Page and Marcotte (1966), nonparametric statistical research continued throughout the 1960s and references to nonparametric tests began to appear in various statistical text books. Page and Marcotte also noted that in the early 1960s, argument against nonparametric statistical use began to occur. Gaito (1960) wrote, "It is encouraging to note that some individuals have been reluctant to embrace wholeheartedly the nonparametric technique" (p. 277). However, Boneau (1961) stated that parametric tests were superior to nonparametric tests, even when the underlying assumptions were violated. Glass, Peckham and Sanders (1972) exemplified the disenchantment with nonparametric statistics, by discouraging its use even when populations were not normal. Their stance was widely accepted and it "encouraged researchers to abandon a highly useful statistical tool" (Blair, 1981, p. 499). Thus began a decline in the popularity of nonparametric statistical techniques.

As outlined in Chapter One, three reasons for the decline were given by Anderson (1961) and Blair (1981), and summarized by Sawilowsky (1990). Each of these three reasons are discussed in detail below. However, it should be noted that Meddis (1984) commented, "We might reasonably assume that the same popularity (of parametric factorial analysis) will one day extend to factorial analysis in the nonparametric domain" (p. 300).

Robustness

The first reason for the historical decline in nonparametric test use involves robustness, in that it is asserted by some that parametric tests are robust, which eliminates the need for alternative tests. One of the difficulties with robustness is agreement on the definition of robustness. According to Berenson and Levine (1992), "Some test procedures are said to be 'robust' because they are relatively insensitive to slight violations in the assumptions" (p. 552). Harwell (1988) offered additional insight on robustness.

> If departures from the underlying assumptions do not seriously impair the distributional properties of a test, the test is considered to be robust. One framework for examining these assumptions is how well the statistical model (i.e., the test) fits the observed data. A good fit implies the test should control its Type I error rate at nominal levels, whereas a poor fit indicates otherwise. (p. 36)

Most statisticians agree in theory with the definitions offered above, but do not agree on a numerical formula for robustness, i.e., quantitative measurements for slightly insensitive or nominal levels. Bradley (1978) stated,

> Not only is there no generally accepted, and therefore standard, quantitative definition of what constitutes robustness, but, worse, claims of robustness are rarely accompanied by any quantitative indication of what the claimer means by the term. (p. 145)

The way to arrive at a quantitative measurement of robustness is to determine a range of p values for a given alpha level, for which the test would be considered robust (Bradley, 1978). Bradley further offers two criterion for quantitative robustness, one stringent and one liberal. The stringent criterion is $.9\alpha \leq \pi \leq 1.1\alpha$. Therefore, for a nominal .05 alpha level, the p value would range from .045 to .055.For the purposes of this paper, the liberal criterion will be adopted which is defined by Bradley as $.5\alpha \leq \pi \leq 1.5\alpha$, where α refers to the nominal significance level, and π which refers to the actual Type I error rate. Therefore, for a nominal .05 alpha level, the p value would range from .025 to .075. For a nominal .01 alpha level, the p value would range from .005 to .015.

Type I Error

Olson (1976) noted, "deficient power merely makes a test less useful, whereas substantial inflation of the Type I error rate makes it positively dangerous" (p. 583). Glass, Peckham and Sanders (1972) stated, "The relevant question is not whether ANOVA assumptions are met exactly, but rather whether the plausible violations of the assumptions have serious consequences on the validity of probability statements based on the standard assumptions" (p. 507). Glass, et al. (1972) presented many findings which support their argument that there are few negative consequences with regard to Type I (and for that matter Type II error) for violation of normality when performing ANOVA or ANCOVA.

Bradley (1978) examined the t test and the F test for robustness in situations where parametric assumptions were violated and found the tests did not always meet the robustness criterion. He found his conclusions "intriguing since these tests are often portrayed as being virtually immune to violation of the normality assumption" (p. 147). Micceri (1989) advised, "considerable research suggests that parametric statistics frequently exhibit either relative or absolute nonrobustness in the presence of certain nonnormal distributions" (p. 157).

One suggestion for addressing this disagreement was offered by Harwell (1988), as follows.

> In general, researchers should not automatically rely on the robustness of any PAR test; some are robust and some are not. The mediating factor in simulation evidence with respect to Type I error control seems to be sample size. Equal or large sample sizes go a long way toward making PAR tests robust, whereas small and unequal samples go a long way toward making these tests sensitive to assumption violations. (p. 36)

Type II Error

"Thus the robustness issue is related not only to Type I error, but also to Type II error, the complement of the power of a statistical test" (Sawilowsky, 1990, p. 98). "Specifically, a test is called robust when its significance level (Type I error probability) and power (one minus Type II error probability) are insensitive to departures from the assumptions on which it is derived" (Ito, 1980, p. 199).

Power is defined as, "The probability of rejecting the null hypothesis when it is in fact false" (Runyon & Haber, 1991, p. 452). Power can be viewed as one minus the probability of a Type II error (Runyon & Haber, 1991). The formula is, $\underline{P} = 1 - \beta$, where β is equal to the actual Type II error probability (Ito, 1980, p. 209, Still & White, 1981, p. 244).

Because Type II error is accepting a false null hypothesis, this only occurs when the null hypothesis is false. Type II error is the degree to which the statistic rejects the null hypothesis when it is false, when compared to normality under non-normality. Scheffe' (1959) advised that even if, for example, the F test shows fairly high power under a particular distribution, there may be another test that is more powerful than the F test under the same distribution.

Power is a function of several factors. The first is sample size, the second is the alpha level, and the third is the nature of the alternative hypothesis. (Runyon & Haber, 1991).

Comparative Power

As was noted above, the second reason for the historical decline in nonparametric test use, is that "it is assumed that nonparametric tests are less powerful than their parametric counterparts, apparently regardless of the shape of the population from which the data were sampled" (Sawilowsky, 1990, p. 92).

Comparative power is related directly to the purpose of this study in that a comparison will be made of several alternatives to the F test (parametric test) in order to determine, (a) which of the three nonparametric alternatives chosen for this study is the most powerful and, (b) if any of the nonparametric alternatives are more powerful than the F test under various distributions, sample sizes and alpha levels.

In discussing the issue of parametric vs nonparametric tests and the associated power of each, Harwell (1988) wrote, "here simulation evidence and conventional wisdom collide" (p. 36). As with robustness, there is disagreement in this area.

One thing that is generally agreed on is that when dealing with a normal distribution, very little power is lost when using nonparametric statistics rather than parametric statistics. Blair and Higgins (1980) showed that the nonparametric Wilcoxon Rank Sum exhibited more power than the parametric Student's $\underline{t}$ test when calculated on samples drawn from a common nonnormal distribution. In a comparison of the Wilcoxon Rank Sum, Mann-Whitney U, Student's $\underline{t}$, and alternate $\underline{t}$ tests for means of normal distributions, Gibbons and Chakraborti (1991) concluded, "we compared the power of these tests and found very little difference. This means that very little power will be lost if the Mann-Whitney U test is used instead of tests that require the assumption of normal distributions" (p. 258). The area where the controversy exists, however, is when the distribution is nonnormal. Harwell (1988) stated,

> For example, simulations have shown that for a variety of non-normal, uni-modal distributions often observed in practice, the power advantages of NPAR over PAR tests can be greater than 20

> points. In other words, an NPAR test under these circumstances would, over the long run, reject a false null hypothesis 20% of the time more often than a PAR competitor. (p. 36)

On the other hand, Rasmussen (1985) reported, "In contrast to the Blair and Higgins (1980) study, a parametric test--when corrected for outliers--shows superior power to the non-parametric test" (p. 509). Rasmussen used mixed-normal samples for his study.

In review of the disagreement between Rasmussen and Blair and Higgins, Sawilowsky (1990) discredited Rasmussen's study by advising that Rasmussen gave unequal treatment to each statistic and limited the distributions to functions of the normal curve. Sawilowsky (1990) concluded, "Therefore, the findings should not be generalized to distributions that are not so amenable to transformation to normality" (p. 100).

Asymptotic Relative Efficiency

Iman, Hora and Conover (1984) described the Asymptotic Relative Efficiency (ARE) in a two-sample layout, for example, as, "the ARE of a test is related to its power, which means there are settings where the nonparametric tests have more power than the paired t test as well as cases where they have less power" (p. 674). The ARE is also known as Pitman efficiency, and is a rough predictor of power to be used when the sample sizes are large. The ARE is widely used because it can compare tests under standardized conditions (Bradley, 1968).

Although a direct correlation has not been found, it has been determined that the greater the distance of the ARE from 1, the increasingly more or less powerful the test. For example, as the ratio becomes increasingly less than one, the nonparametric test is increasingly less powerful than the parametric test. Conversely, as the ratio becomes increasingly greater than one, the nonparametric test is increasingly more powerful than the parametric test.

Lack of Nonparametric Tests

Finally, the third reason for the historical decline in use of nonparametric procedures has been the lack of nonparametric tests, particularly, for more complex research designs (Harwell, 1988; Sawilowsky, 1990). Two related issues under this category have been "the absence of serious coverage of NPAR tests in many graduate training programs in psychology and education" (Harwell, 1988, p. 35), and "the perception that, by comparison, relative few nonparametric tests are available in statistical computing programs (e.g., rank regression)" (Harwell, 1990, p. 144).

Nonparametric Alternatives to ANOVA

Two of the most common nonparametric alternatives to ANOVA are the Kruskal-Wallis and Friedman tests (Ingram & Monks, 1992). The Kruskal-Wallis test is a nonparametric alternative to one-way ANOVA. The assumptions underlying this test are that the "samples for each group are independent one from another and are randomly drawn from their respective populations. Also, the observations are required to be ordinal or higher scale" (Ingram & Monks, 1992, p. 808). Berenson and Levine (1992) stated, "the Kruskal-Wallis procedure has proven to be almost as powerful as the F test under conditions appropriate to the latter and even more powerful than the classical procedure when its assumptions are violated" (p. 574). Hodges and Lehmann (1956) determined that when compared to ANOVA, the ARE of the Kruskal-Wallis test may be as high as infinity, and has a lower limit of .864.

The Friedman test is a nonparametric alternative to two-way ANOVA. The assumption is, "Samples are dependent under all levels" (Ingram & Monks, 1992, p. 827). "The Friedman test is essentially a two-way analysis of variance (ANOVA) on ranked data" (Berenson & Levine, 1992, p. 579).

These tests, by themselves, are appropriate as long as an interaction effect is not present in the experiment. However, interaction effects are fairly common in applied disciplines (Tate & Clelland, 1957). Toothaker and Chang (1980) noted, "suitable tests for main effects and interactions were needed to complete the package of those methods not making the normality assumption" (p. 169). In discussing nonparametric tests, Klugh (1986) stated, "Unfortunately they are not as powerful as the $\underline{t}$ test and the F test and, of course, none of them can reveal the subtle interaction effects for which the ANOVA is so well-suited" (p. 376).

Nonparametric Tests of Interaction

Anderson (1961) noted the lack of nonparametric tests of interaction.

> Inferences based on nonparametric tests of interaction would presumably be less sensitive to certain types of scale changes. However, caution would still be needed in the interpretation....The problem is largely academic, however, since few nonparametric tests of interaction exist. (p. 314)

However, over the last 25 years, many new nonparametric tests for interaction have been developed. Sawilowsky (1990) listed ten nonparametric tests for interaction, most of which were developed during the 1980s. Each one is briefly summarized below. More detailed definitions are given for the last three tests for interaction: Bradley's Collapsed and Reduced Technique, Harwell and Serlin's L test, and Blair and Sawilowsky's Adjusted Rank Transform test, as these three nonparametric tests for interaction have been explored using Monte Carlo techniques in this comparison study.

Approximate Randomization Test

The Approximate Randomization test was developed by Still and White (1981). It is a permutation version that has been shown to work well. However, sophisticated computers are required to perform the test. Still and White performed a Monte Carlo study and were pleased with the test's performance, but recognized the computer-related drawbacks. They concluded of their test, "It is therefore preferable to the parametric test where suitable computing facilities are available and computing time is not at a premium" (p. 252). During the last three years, the increasing sophistication of computers has made the permutation test more feasible for future use and further studies examining its performance.

Moment Approximation Test

Developed by Berry and Mielke (1983), the Moment Approximation Test is based on calculus, is difficult to perform, and cannot be used for data distributions that have undefined higher moments, such as the Cauchy distribution.

Rank Transform Test

The steps involved in performing rank transformation were outlined by Conover and Iman (1981). "Simply replace the data with their ranks, then apply the usual parametric $\underline{t}$ test, F test, and so forth, to the ranks. We call this the rank transformation (RT) approach" (p. 124).

Conover and Iman (1981) found the RT effective in detecting interactions in some circumstances. Sawilowsky (1985) performed a Monte Carlo study which showed the Rank Transform test was invalid in the 2 x 2 x 2 layout. Also, Blair and Higgins (1985) demonstrated the Rank Transform is invalid in the context of the dependent sample $\underline{t}$ test.

Random Normal and Expected Scores Transform Test

The Random Normal and Expected Scores Transform test is a modification of the Rank Transform test described above, in which the original data is ranked, and the ranks are replaced by expected normal scores which are dependent on the sample size $\underline{n}$. Sawilowsky (1989) found that although this procedure outperformed the Rank Transform test, it still was not as powerful as ANOVA, especially with small sample sizes.

Other Aligned Ranks-Based Tests

Aligned rank procedures are "based on data which is first aligned (adjusted for row and column effects) and then ranked" (Iman & Conover, 1976, p. 8). In order to obtain the interaction, "each observation is 'aligned' by subtracting the row mean and the column mean for the row and column that contain that observation. The residuals are then ranked on an overall basis, ... and the usual F statistic for interaction is computed on these ranks" (Iman & Conover, 1976, p. 8).

Hettmansperger (1984) proposed a procedure for testing interaction effects which involved aligning the effects not being tested. These aligned effects were then ranked and a statistic was computed for the aligned ranks. Harwell (1991) noted the tediousness of this procedure.

Hodges and Lehmann (1962), suggested a test which uses a chi-square value based on (a-1)(b-1) degrees of freedom to test the hypothesis of no interaction.

Studies of the various aligned ranks-based tests found that these tests compared favorably with ANOVA, and in some cases performed better than ANOVA. (For more information, see Conover & Iman, 1981; Groggel, 1987; and Marascuilo & McSweeney, 1977).

Extended Median Test

Wilson (1956) presented a median test for ANOVA designs. This test was expanded by Shoemaker (1986), and is based on a proportion of counts in cells which are above or below the median.

This test was found to be robust with small sample sizes and in certain circumstances, a good alternative to ANOVA (Shoemaker, 1986).

Large- and Small-Samples Patel and Hoel Test

The original Large- and Small-Samples test developed by Patel and Hoel (1973) was difficult to compute. Further studies examining power and robustness have not been done at this time.

Bradley's Collapsed and Reduced Technique

Bradley (1979) introduced a Collapsed and Reduced Technique to test for interaction. Bradley's approach involved the following three steps.

> First, the data are collapsed over all variables except replicates and those variables whose interaction is to be tested. Then, by a series of subtractions, the resulting data table is reduced to one in which all lower-order effects have been eliminated and which now contains only as many cells as there are degrees of freedom for the interaction of interest. At this point, a test for main effects applied to the cells of the reduced table is equivalent to a test for interaction in the original data base. So the final step is to apply the appropriate nonparametric test for main effects to the data in the reduced table. (p. 177).

Bradley (1968) provided several limitations in that the design must be balanced (the numbers of observations per cell must be equal) and the measurement may not yield many tied observations. In addition, the researcher must be aware that the placement of the data into the matrix may be purposely manipulated to yield biased results. Therefore, the researcher must avoid situations in which the data is rearranged within the matrix until the desired results are obtained.

A drawback of the test is that it is difficult to replicate because it is conceivable that the data may be placed in various matrix designs, thereby obtaining differing probability outcomes and possibly leading to different results.

Harwell and Serlin's L Test

Puri (1969) and Puri and Sen (1971) produced a test they called an L statistic, which involved ranking data and then computing a pure- or mixed-rank form L statistic. Harwell (1991) and Harwell and Serlin (1989) conducted several simulation studies on a form of the Puri and Sen L statistic. In 1990, Harwell noted the shortage of computer software for nonparametric statistics, and consequently adapted Puri and Sen's L statistic in trace criterion form for use with existing computer programs.

Harwell outlined the steps of the test by first ranking the data, and then using a packaged computer program to calculate, (a) simple regression; (b) multiple regression; (c) fixed-effects, one-way ANOVA; (d) fixed-effects, one-way ANCOVA; (e) fixed-effects, one-way Multiple Analysis of Variance (MANOVA); and (f) canonical correlation on the values. Finally, Harwell calculated the L statistic on the computerized results. Harwell (1990) concluded that, "for a number of nonnormal distributions, the power of L exceeds that of its normal-theory counterparts, but for small samples this power may be undesirably low" (p. 147).

The interaction effect tested by Harwell (1991) is the same as a test presented by

Meddis (1984, pp. 299-313), which uses a statistic similar to the Kruskal-Wallis test.

The requirements for using Puri and Sen's test are that, "scores are sampled from a common distribution, are independent across subjects, and that the sample size is large enough to ensure the validity of the chi-square approximation" (Harwell, 1990, 155).

Blair and Sawilowsky's Adjusted Rank Transform Test

Blair and Sawilowsky (1990) proposed an adjustment to the standard rank transform test, and summarized its use in an A x B layout.

> Subtract the mean (or a trimmed mean, winsorized mean, or robust mean) of the column from the original observations in each column. Then, subtract the mean of the row from the observations of each row. This removes the presence of the two main effects, leaving only the effect due to the interaction. Of course, the researcher never knows what the real main effects are, but subtracting the mean is the best unbiased guess over the long run. The resultant values are pooled together and ranked, and the ranks are returned to their respective cells. Then, the usual parametric test is performed. (p. 3)

(A test developed independently by Fawcett and Salter, 1984, is computed in a similar fashion except that in addition to the row and column means, the grand mean is also subtracted from the observations.)

When compared to the ANOVA, Blair and Sawilowsky (1990) determined this form of the Adjusted Rank Transform test to have "superior power properties" (p. 4). However, the study expressed the need to compare the Adjusted Rank Transform test with other nonparametric tests before recommending use of the test.

In a soon to be published study, Higgins and Tashtoush examined an aligned rank transform test for interaction which is similar to both Blair and Sawilowsky's and Fawcett and Salter's test. When the normality assumption was violated, Higgins and

Tashtoush found the test valid for small and moderate sample sizes under certain circumstances.

CHAPTER THREE

METHODOLOGY

Monte Carlo Study

Many times in research, questions exist which require an answer by an empirical study. A common way to explore the power and robustness of one or more statistical tests is to employ a simulation study using both normal and nonnormal distributions. The most widely used method is a Monte Carlo (MC) study, which is a computer simulation (Harwell, 1990). Harwell (1990) summarized the Monte Carlo process as it relates to violations of underlying assumptions (as is the case in this study).

> In the typical MC study of a given statistical test the following process is repeated for a large number of samples: data are simulated which reflect a specified relationship among variables (but which do not usually conform to the assumptions required for correct application of the test), the statistical test is computed for the data, and the value of the statistical test is recorded. The collection of values of the statistical test provide information on its properties (e.g., the proportion of "significant" values of the test). If the underlying assumptions of the test were satisfied, exact statistical theory would guarantee that the test would have a specified type I error rate and would permit the probability of rejecting a false statistical hypothesis to be computed; MC studies permit these characteristics to be examined when underlying assumptions are violated. (p. 4)

Monte Carlo techniques were used for this study using Unix-based Hewlett Packard Apollo 9000/Series 735 with Hewlett Packard Fortran 77.

Method

The Type I error and power properties of the Analysis of Variance (ANOVA) test for interaction were compared to the following nonparametric alternatives: (a) Bradley's Collapsed and Reduced Technique, (b) Harwell and Serlin's L test, and (c)

Blair and Sawilowsky's Adjusted Rank Transform test. The tests were examined in the context of a 2 x 2 x 2 layout. Because there are 27 adjustments and tests necessary to fully examine the Adjusted Rank Transform test in the context of a 2 x 2 x 2 layout, only the following four, which are a reasonable subset, were considered: a) test of A x B x C, b) test of A x B, c) test of A x C, and d) test of B x C, when all effects are nonnull (see treatment condition (g) below). The data generation model used is as follows.

$$x_{ijkl} = \mu + A_i + B_j + C_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} + e_{ijkl}$$

$i = 1, 2$

$j = 1, 2$

$k = 1, 2$

$l = 7, 21, \text{or } 35$

Both interaction and main effects were modeled. The A main effects were created by making a_1 equal to c and a_2 equal to -c and, where c assumed the values 0.20σ, 0.40σ, 0.60σ, 0.80σ, 1.00σ, and 1.20σ (σ represents the standard deviation of the sampled population). The B and C main effects were modeled in a corresponding way.

The A x B interactions were modeled by making $(ab)_{11} = (ab)_{22} = c$, during which time $(ab)_{21} = (ab)_{12} = -c$, where c assumed the values 0.20σ, 0.40σ, 0.60σ, 0.80σ, 1.00σ, and 1.20. A x C and B x C interactions were modeled in a corresponding way.

The A x B x C interaction was modeled by making $(abc)_{111} = (abc)_{221} = (abc)_{212}$

$= (abc)_{122} = c$, and $(abc)_{211} = (abc)_{121} = (abc)_{112} = (abc)_{222} = -c$, where c assumed the values 0.20σ, 0.40σ, 0.60σ, 0.80σ, 1.00σ, and 1.20σ.

Samples in the sizes discussed below were chosen from the following four theoretical distributions and two real data sets: (a) Gaussian, (b) uniform, (c) t with 3 degrees of freedom, (d) exponential, (e) multi-modal lumpy, and (f) discrete mass at zero. As discussed above, constants were added to create effects and the suitable F statistics were determined. Critical values were then used from the F table to determine the proportions of hypotheses to be rejected for each effect at the .05 and .01 alpha levels. The procedure was repeated on the ranks of the original data and comparisons were made.

Sample Size, Nominal Alpha, and Repetitions

All comparisons involved groups (2 x 2 x 2) of equal sample size. Sample sizes of cell n=7, 21 and 35 were used. Sample size of cell n beginning with seven was chosen because published tables of critical values for the nonparametric statistic provide alpha levels of approximately .05 beginning at cell n=7. Sample size of cell n=21 was chosen to provide a medium-size sample, and cell n=35 was chosen to provide a large-size sample. For both Type I error and power (discussed below) this study's nominal alpha for the parametric statistic was set at .05 and .01. The number of repetitions for the Monte Carlo study was set at 20,000 per experiment.

Generation of Data

Pseudo-random data was generated for the mathematical curves (e_{ijkl} in the model above) by using International Mathematical and Statistical Libraries (IMSL, 1987) subroutines RNNOA and RNUN for Gaussian and Uniform data respectively, RANGEN

(1987) subroutine T1 for the *t* distribution and RNEXP for the exponential distribution. Real data sets from Micceri (1989) were also used as discussed below. A total of four theoretical distributions and two real data sets were used for this study because of the variation in skew and tail weights provided by the different distributions. Each distribution and data set is discussed below.

Gaussian

The Gaussian, or normal distribution was necessary in order to compare nonparametric procedures with parametric tests when the distribution assumption is met. It is also used to verify the accuracy of the algorhythms in the Fortran program.

The functional form is,

$$f(x) = \frac{1}{\sqrt{2}\,\pi} e^{-x/2}$$

Uniform

The uniform distribution was selected to provide data from a light-tailed (when compared to the Gaussian) distribution.

The functional form is,

$$f(x) = 1 \quad , \text{where } 0 \leq x \leq 1$$

t (df = 3)

This distribution is the *t* distribution with three degrees of freedom. It was chosen to provide a heavy-tailed distribution.

The functional form is,

$$f(x) = \frac{\tau(2)}{\tau(2/3)\sqrt{3\pi}\,(1+x^2/3)^2}$$

Exponential

This distribution is asymmetrical and heavy-tailed. It was chosen to provide a skewed distribution. The functional form is,

$$f(x) = e^{-(x-\mu)}, \text{ where } x >= \mu$$

Multi-modal Lumpy

The Multi-modal Lumpy is a real data set which is furnished by Micceri (1989). Sawilowsky and Blair (1992) provide a graph of the distribution in their Figure 6 (p. 355). This distribution is based on a set of achievement test scores with a mean = 21.15, median = 18, standard deviation = 11.9, skew = .19 and kurtosis (scaled so that normal kurtosis is 3.0) = 1.8.

Discrete Mass At Zero

This real data set (Micceri, 1989) is taken from psychometric scores. It has the unusual feature of the majority of scores being zero with a gap to the next valid score ranging from 8 through 11, and is depicted in Sawilowsky and Blair's (1992) Figure 2 (p. 354). This distribution has a mean = 1.85, median = 0, standard deviation = 3.8, skew = 1.65 and kurtosis = 3.98.

Type I Error and Power

The Type I error rates and power of all tests were analyzed under the circumstances listed. The effects are statistically equivalent according to normal theory test (Shaffer, 1977) and therefore the fundamental concern is number of nonnull effects rather than which effects are nonnull. The condition relating to Type I error is: all effects null (c = 0).

The conditions relating to power are: (a) all effects null except for the A main effect which is nonnull; (b) all effects null except for the A and B main effects which are nonnull; (c) all effects are null except for the A, B, and C main effects which are nonnull; (d) all effects are null except for the A, B and C main effects and the A x B interaction which are nonnull; (e) all effects are null except for the A, B, and C main effects and the A x B and A x C interactions which are nonnull; (f) all effects are null except for the A, B and C main effects and the A x B, A x C and B x C interactions which are nonnull; (g) the A, B and C main effects and the A x B, A x C, B x C and A x B x C interactions are nonnull.

Worked Examples of Three Nonparametric Tests

Bradley's Collapsed and Reduced Technique, Harwell and Serlin's L test, and Blair and Sawilowsky's Adjusted Rank Transform test, which are the three non-parametric tests chosen for examination in this dissertation, were performed with a fabricated data set in order to show the step-by-step process for each of these tests. A summary of the Analysis of Variance (ANOVA) results is also shown; however, the process for ANOVA is not delineated because of the prevalence of ANOVA in computerized statistical packages which are available for verification.

Dixon and Massey (1969) presented an interaction effect found in education. In the example, two groups of people, (Men and Women) and two methods of teaching (Teaching method A and B) were examined as independent variables. The dependent variable was a final examination score for each of the individuals. Dixon and Massey's example used only twelve individuals, which is too small to permit the use of some of the statistical tests used in this study; therefore, Dixon and Massey's example has been expanded to twenty-four observations by creating additional scores. Although this dissertation examines the statistical tests in the context of a 2 x 2 x 2 layout, a 2 x 2 layout was chosen for the worked examples to simplify the demonstration test calculation.

Table 1
Fabricated data set collected on students, with method of teaching and gender as the two independent variables

	Teaching method A (B1)	Teaching method B (B2)
Men (A1)	4	4
	4	3
	7	7
	5	5
	8	5
	9	6
Women (A2)	10	9
	12	8
	13	5
	11	7
	14	8
	15	6

Bradley's Collapsed and Reduced Technique

Bradley's test collapses and reduces the data into one table. When using a 2 x 2 x 2 layout, it is necessary to collapse the data over the third main effect. However, in the 2 x 2 layout shown here, this step is eliminated and the first step is to reduce the rows by subtracting the first score in the first cell of A1 (Men) from the first score in the first cell of A2 (Women). This process is repeated for each observation, as shown in Table 2.

Table 2
Step 1 of Bradley's Collapsed and Reduced Technique

	Teaching method A (B1)	Teaching method B (B2)
Men - Women	4 - 10 = -6	4 - 9 = -5
(A1-A2)	4 - 12 = -8	3 - 8 = -5
	7 - 13 = -6	7 - 5 = 2
	5 - 11 = -6	5 - 7 = -2
	8 - 14 = -6	5 - 8 = -3
	9 - 15 = -6	6 - 6 = 0

The second step of Bradley's test involves further subtraction; however, this time the subtraction is across columns, which will reduce the two columns (Teaching method A and B) into one column, as shown in Table 3.

Table 3
Step 2 of Bradley's Collapsed and Reduced Technique

	Teaching method A - Teaching method B (B1-B2)
Men and Women	-6 - (-5) = -1
(A1-A2)	-8 - (-5) = -3
	-6 - 2 = -8
	-6 - (-2) = -4
	-6 - (-3) = -3
	-6 - 0 = -6

The third step of Bradley's test involves ranking the data and performing the appropriate nonparametric test on the ranks. In this case, Bradley recommends performing a Wilcoxon Sign-Rank test on the ranks. The ranks are shown in Table 4.

Table 4
<u>Step 3 of Bradley's Collapsed and Reduced Technique</u>

Reduced Scores	Ranks (Sign Retained)
-1	-6
-3	-4.5
-8	-1
-4	-3
-3	-4.5
-6	-2

The test statistic is obtained by summing the ranks which are positive (Berenson & Levine, 1992). Because there are no positive ranks, the sum of the positive ranks is 0. Thus, the observed value of 0 is compared to the critical values in the Wilcoxon table for $\underline{n} = 6$ at the .05 alpha level, which are 2 (lower tail) and 19 (upper tail). If the obtained value is less than the lower-tailed critical value or greater than the upper-tailed critical value, the null hypothesis is rejected indicating the interaction is significant. In this example, Bradley's Collapsed and Reduced Technique shows the interaction to be significant.

Harwell and Serlin's L Test

The first step in this nonparametric test is to rank the original scores (Table 5).

Table 5
Step one of Harwell and Serlin's L Test

	Teaching method A (B1)	Rank	Teaching method B (B2)	Rank
Men (A1)	4	3	4	3
	4	3	3	1
	7	12	7	12
	5	6.5	5	6.5
	8	15	5	6.5
	9	17.5	6	9.5
Women (A2)	10	19	9	17.5
	12	21	8	15
	13	22	5	6.5
	11	20	7	12
	14	23	8	15
	15	24	6	9.5

The sum of squares is then computed on the ranks. A computerized statistical package may be used to compute the sum of squares. Table 5 shows the results of the sum of squares performed on the ranks.

Table 6
Sum of squares on the ranks for Harwell and Serlin's L Test

Source of variation	Sum of squares
Gender	495.042
Method	216.000
Interaction	51.042
Total	1138.000

The next step is to use the sum of squares for the interaction (ABSS) and sum of squares total (TSS) to calculate the statistic in trace criterion form, which is computed using the formula below, and where $\underline{N}$ is equal to the number of observations in the data set.

$$L = (\underline{N} - 1) \times ABSS/TSS$$

$$L = (24-1) \times 51.042/1138.000 = 1.0316$$

According to this formula, the obtained value is 1.0316.

A chi-square table is then used to determine the critical value. The degrees of freedom are found by subtracting one from each of the independent variables, and multiplying the values together (A - 1) x (B - 1), as follows:

$$(2 - 1) \times (2 - 1) = 1$$

Therefore, at a .05 alpha level, the critical value is 3.841. Because the obtained value is less than the critical value, the interaction is not significant according to Harwell and Serlin's L test.

Blair and Sawilowsky's Adjusted Rank Transform Test

The first step of Blair and Sawilowsky's Adjusted Rank Transform test is to calculate the row and column means, after which both means are subtracted from each score, as shown in Table 6. The mean of A1 is 5.58, the mean of A2 is 9.83, the mean of B1 is 9.33, and the mean of B2 is 6.08.

Table 7
Subtraction of the row and column means for Blair and Sawilowsky's Adjusted Rank Transform Test

	Teaching method A (B1)	Teaching method B (B2)
Men (A1)	4-5.58-9.33 = -10.91	4-5.58-6.08 = -7.66
	4-5.58-9.33 = -10.91	3-5.58-6.08 = -8.66
	7-5.58-9.33 = -7.91	7-5.58-6.08 = -4.66
	5-5.58-9.33 = -9.91	5-5.58-6.08 = -6.66
	8-5.58-9.33 = -6.91	5-5.58-6.08 = -6.66
	9-5.58-9.33 = -5.91	6-5.58-6.08 = -5.66
Women (A2)	10-9.83-9.33 = -9.16	9-9.83-6.08 = -6.91
	12-9.83-9.33 = -7.16	8-9.83-6.08 = -7.91
	13-9.83-9.33 = -6.16	5-9.83-6.08 = -10.91
	11-9.83-9.33 = -8.16	7-9.83-6.08 = -8.91
	14-9.83-9.33 = -5.16	8-9.83-6.08 = -7.91
	15-9.83-9.33 = -4.16	6-9.83-6.08 = -9.91

The next step is to rank the data, as shown in Table 7.

Table 8
Ranked data for Blair and Sawilowsky's Adjusted Rank Transform Test

	Teaching method A (B1)	Rank	Teaching method B (B2)	Rank
Men (A1)	-10.91	2.5	-7.66	13
	-10.91	2.5	-8.66	8
	-7.91	12	-4.66	23
	-9.91	5	-6.66	17.5
	-6.91	16	-6.66	17.5
	-5.91	20	-5.66	21
Women (A2)	-9.16	6	-6.92	15
	-7.16	14	-7.92	10.5
	-6.16	19	-10.92	1
	-8.16	9	-8.92	7
	-5.16	22	-7.92	10.5
	-4.16	24	-9.92	4

After the adjusted scores are ranked, ANOVA is calculated on the ranks (Table 8), which shows the interaction significant at a .05 alpha level.

Table 9
ANOVA calculated on the adjusted ranks

Source of variation	Sum of squares	DF	Mean square	F	Sig. F
Gender	n/a				
Method	n/a				
Interaction	322.667	1	322.667	7.894*	.011
Residual	817.500	20	40.875		
Total	1144.500	23	49.761		

*$p < .05$.

Analysis of Variance (ANOVA)

The ANOVA results are shown in Table 9. These results were obtained using SPSS/PC+ (version 4.0 for DOS). Table 9 shows that the interaction is significant at a .05 alpha level.

Table 10
ANOVA calculated on the original scores

Source of variation	Sum of squares	DF	Mean square	F	Sig. F
Gender	108.375	1	108.375	35.436*	.000
Method	63.375	1	63.375	20.722*	.000
Interaction	26.042	1	26.042	8.515*	.009
Residual	61.167	20	3.058		
Total	258.958	23	11.259		

*$p < .05$.

CHAPTER FOUR

RESULTS AND DISCUSSION

The results from the computer program were summarized into both tables and figures (graphs) and are not included in this dissertation in their entirety because they number over 1,500 pages. In order to conserve space, typical examples are presented in the tables. The remaining results will be archived in ERIC. The results which appear in the following tables and figures were conducted at the .05 alpha level.

Table 11

Table 11 was developed to show the average alpha levels of three of the four tests under the various distributions/data sets when there was no treatment, i.e., all effects were null. Alpha levels were not calculated for Blair and Sawilowsky's Adjusted Rank Transform test because the test is not to be used when all the effects are null. Two of the three sample sizes studied are displayed in this table ($\underline{n}=7$ and $\underline{n}=35$). The "Statistic" column indicates which test statistic was used. "F" indicates the Analysis of Variance results, "B" indicates Bradley's Collapsed and Reduced Technique results, and "H-S" indicates Harwell and Serlin's L test results. The "Population" column indicates which distribution/data set was used.

Tables 12 - 23

Tables 12 through 23 show the alpha levels (α) and the power (1-ß) for various effect sizes, statistical tests, models and sample sizes. Each table provides results for one of each of the six distributions/data sets, as indicated in the title. Each table also represents results from either a sample size of $\underline{n}=7$ or $\underline{n}=35$. The "Outcome" column indicates whether the recorded result represents Type I error (α) or power (1-ß). The "Stat" column indicates which test statistic was used. "F" indicates the Analysis of Variance results (referred to as ANOVA in the following discussion), "B" indicates Bradley's Collapsed and Reduced Technique results (referred to as Bradley in the following discussion), "H-S" indicates Harwell and Serlin's L test results (referred to as Harwell-Serlin in the following discussion), and "B-S" indicates Blair and Sawilowsky's Adjusted Rank Transform test results (referred to as Blair-Sawilowsky in the following discussion). The "c" column stands for the value of the constant which was added as a treatment to produce the nonnull effects. These tables show the results for the constants .2 and .8. Additional levels of .4, .6, 1.0 and 1.2 were also examined and produced similar results.

The "Number of nonnull effects" column indicates how many and which effects were modeled. Column 1 indicates that all effects were null except for the A main effect which was nonnull. Column 2 indicates that all effects were null except for the A and B main effects which were nonnull. Column 3 indicates that all effects were null except for the A, B and C effects which were nonnull. Column 4 indicates that all effects were null except for the A, B and C main effects and the B x C interaction which was nonnull. Column 5 indicates that all effects were null except for the A, B, and C main effects and the B x C and A x C interactions which were nonnull. Column 6 indicates that all effects were null except for the A, B and C main effects and the B x C, A x C, and A x B interactions which were nonnull. Column 7 indicates that the A, B and C main effects and the B x C, A x C, A x B and A x B x C interactions were nonnull.

In most cases, the alpha and power levels represent averages. For instance, the alpha level for the F test under Column 1 was calculated by summing all of the null effects and dividing by the number of null effects found within that column. Because Column 1 had an A main effect only, the null effects were the B and C main effects and the A x B, A x C, A x C, and A x B x C interactions. Therefore, the alpha levels for the B and C main effects and the A x B, A x C, B x C, and A x B x C interactions were summed, then divided by 6. This procedure was repeated throughout the models with several exceptions. In some cases, there was only one null or nonnull effect which could not be averaged, and were therefore singularly recorded. For example, Column 6, has just one null effect (A x B X C interaction). In that case, the A x B x C interaction alpha level was used.

The other exception to averaging were the results from the Blair-Sawilowsky test, which removed each previous nonnull effect, and produced a power level for the newest effect only. For example, Column 3 had an A, B and C main effect. The A main effect was introduced in Column 1, the B main effect in Column 2, and the C main effect was introduced in Column 3. Both the F test and Harwell-Serlin test produced power levels for all three main effects. The Blair-Sawilowsky test removed the A and B main effect and produced a power level for the C main effect only. Therefore, the Blair-Sawilowsky power levels were not averaged.

The nonparametric tests chosen for this dissertation were specifically designed to test for interactions. In order to provide a thorough study, these tests were also examined when only main effects were present. There are several instances when these nonparametric tests cannot be used. These cases are explained in the following two paragraphs.

The alpha and power levels which appear under the A main effect for Blair-Sawilowsky were calculated by using an unadjusted Rank Transform (RT) test. For the purposes of this dissertation when testing the first treatment model, it would not make sense to use the Adjusted Rank Transform. When there is only one nonnull effect there is nothing left to test after the adjustment; therefore the Adjusted Rank Transform was not used in the first treatment model, and an unadjusted Rank Transform was used in its place. For example, if the only effect is an A main effect, and it is removed in the adjustment, the effect is no longer detectable.

Although the A main effect was the first effect modeled in this dissertation, it is possible for example, that a C main effect may be the only nonnull effect in a data set. In an actual data analysis it is possible to use the Adjusted Rank Transform to test for a singular main effect by adjusting everything but the singular main effect. For example, if the location and number of effects are unknown, all but one effect would be systematically removed through adjustment. The results would be examined, and then all but the next effect would be removed through adjustment. This procedure would be repeated until all of the possible effects had been examined.

There are no power levels recorded in Columns 1 through 3 for the Bradley test because it is specifically designed to be used for testing interactions, which were not introduced until Column 4.

Figures 1 - 126

Figures 1 through 126 were developed to show the comparative power of the various tests under the .05 alpha level. Because the results for alpha=.010 were similar, they are not shown here. There is a different graph for each of the following conditions:

(a) sample size (n=7, n=21 and n=35), (b) distribution/data set (Gaussian, uniform, t with 3 degrees of freedom, exponential, multi-modal lumpy and discrete mass at zero), and (c) treatment (A main effect; A and B main effects; A, B and C main effects; A, B, and C main effects and B x C interaction; A, B and C main effects and B x C and A x C interactions; A, B and C main effects and B x C, A x C and A x B interactions; A, B and C main effects and B x C, A x C, A x B and A x B x C interactions).

The "Power" axis ranges from 0 to 1. A power level of 1 indicates that the test is detecting each treatment. Although the axis indicates 1, theoretically, it is impossible to achieve a power of 1. With alpha set at .05, any entry larger than .05 represents the proportion of rejections of a false null hypothesis (or power).

The "Effect Size" axis ranges from .2 to 1.2. The effect size indicates the magnitude of the treatment, which in this study is the percent of a standard deviation of the distribution sampled added to the treated scores. For example, an effect size of .2 indicates that the values of the treated scores were increased by 20% of a standard deviation.

The tables and figures in the remainder of this chapter are represented without further comment.

Table 11
Type I error rates of the ANOVA F, Bradley and Harwell-Serlin tests when all effects are null and sampling is from various population shapes.

Population	Statistic	Sample size 7	Sample size 35
Gaussian	F	0.050	0.050
	B	0.046	0.050
	H-S	0.050	0.050
Uniform	F	0.051	0.051
	B	0.048	0.049
	H-S	0.050	0.050
t (df = 3)	F	0.046	0.049
	B	0.046	0.050
	H-S	0.049	0.051
Exponential	F	0.047	0.049
	B	0.047	0.050
	H-S	0.051	0.050
Multi-modal lumpy	F	0.050	0.049
	B	0.044	0.049
	H-S	0.050	0.049
Discrete mass at zero	F	0.049	0.050
	B	0.046	0.050
	H-S	0.050	0.050

Table 12
Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution and n=7.

c	Stat	Outcome	Number of nonnull effects 1	2	3	4	5	6	7
0.20	F	α	0.052	0.050	0.050	0.050	0.047	0.052	n/a
		1-ß	0.108	0.113	0.114	0.115	0.113	0.113	0.115
	B	α	0.048	0.045	0.048	0.046	0.047	0.047	n/a
		1-ß	n/a	n/a	n/a	0.089	0.085	0.085	0.087
	H-S	α	0.050	0.048	0.048	0.046	0.044	0.046	n/a
		1-ß	0.106	0.108	0.107	0.107	0.102	0.100	0.098
	B-S	α	0.052	0.050	0.051	0.051	0.049	0.053	n/a
		1-ß	0.108	0.110	0.112	0.114	0.111	0.113	0.112
0.80	F	α	0.049	0.048	0.049	0.049	0.049	0.049	n/a
		1-ß	0.833	0.834	0.833	0.834	0.833	0.834	0.835
	B	α	0.046	0.048	0.046	0.044	0.045	0.045	n/a
		1-ß	n/a	n/a	n/a	0.665	0.665	0.665	0.671
	H-S	α	0.034	0.024	0.017	0.013	0.015	0.027	n/a
		1-ß	0.815	0.777	0.719	0.666	0.567	0.451	0.317
	B-S	α	0.049	0.050	0.050	0.049	0.049	0.049	n/a
		1-ß	0.815	0.812	0.815	0.814	0.816	0.813	0.809

$\alpha=.05$

Table 13
Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution and n=35.

c	Stat	Outcome	Number of nonnull effects 1	2	3	4	5	6	7
0.20	F	α	0.050	0.049	0.051	0.050	0.048	0.051	n/a
		1-ß	0.379	0.384	0.387	0.385	0.384	0.384	0.385
	B	α	0.050	0.049	0.050	0.050	0.048	0.050	n/a
		1-ß	n/a	n/a	n/a	0.356	0.358	0.354	0.355
	H-S	α	0.049	0.047	0.048	0.045	0.043	0.043	n/a
		1-ß	0.364	0.366	0.364	0.359	0.351	0.347	0.336
	B-S	α	0.050	0.049	0.051	0.050	0.049	0.051	n/a
		1-ß	0.365	0.370	0.375	0.377	0.376	0.371	0.372
0.80	F	α	0.050	0.050	0.050	0.050	0.051	0.050	n/a
		1-ß	0.999	*					
	B	α	0.050	0.050	0.051	0.050	0.050	0.050	n/a
		1-ß	n/a	n/a	n/a	0.999	*		
	H-S	α	0.035	0.024	0.019	0.021	0.044	0.021	n/a
		1-ß	0.999	0.999	0.999	0.999	0.999	0.998	0.973
	B-S	α	0.050	0.050	0.050	0.051	0.050	0.049	n/a
		1-ß	0.999	*					

*power > .9999

α=.05

Table 14
Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution and n=7.

c	Stat	Outcome	1	2	3	4	5	6	7
			Number of nonnull effects						
0.20	F	α	0.051	0.050	0.051	0.050	0.053	0.050	n/a
		1-ß	0.111	0.113	0.108	0.113	0.113	0.111	0.113
	B	α	0.046	0.046	0.047	0.045	0.047	0.046	n/a
		1-ß	n/a	n/a	n/a	0.085	0.084	0.084	0.086
	H-S	α	0.049	0.047	0.047	0.045	0.047	0.045	n/a
		1-ß	0.108	0.108	0.099	0.102	0.100	0.095	0.094
	B-S	α	0.051	0.050	0.051	0.050	0.052	0.051	n/a
		1-ß	0.108	0.112	0.105	0.107	0.109	0.106	0.108
0.80	F	α	0.052	0.050	0.051	0.051	0.050	0.051	n/a
		1-ß	0.832	0.888	0.833	0.834	0.835	0.836	0.836
	B	α	0.047	0.046	0.049	0.046	0.049	0.047	n/a
		1-ß	n/a	n/a	n/a	0.658	0.661	0.662	0.662
	H-S	α	0.037	0.026	0.019	0.016	0.013	0.019	n/a
		1-ß	0.777	0.730	0.677	0.631	0.548	0.448	0.329
	B-S	α	0.051	0.050	0.051	0.051	0.050	0.051	n/a
		1-ß	0.777	0.785	0.777	0.772	0.774	0.773	0.769

α=.05

Table 15
Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution and n=35.

c	Stat	Outcome	Number of nonnull effects 1	2	3	4	5	6	7
0.20	F	α	0.050	0.050	0.050	0.051	0.049	0.050	n/a
		1-ß	0.380	0.383	0.385	0.382	0.385	0.384	0.382
	B	α	0.049	0.049	0.049	0.051	0.048	0.050	n/a
		1-ß	n/a	n/a	n/a	0.344	0.349	0.349	0.351
	H-S	α	0.049	0.048	0.047	0.046	0.044	0.046	n/a
		1-ß	0.365	0.358	0.355	0.343	0.335	0.325	0.309
	B-S	α	0.050	0.050	0.050	0.050	0.050	0.050	n/a
		1-ß	0.366	0.364	0.365	0.358	0.355	0.356	0.344
0.80	F	α	0.050	0.050	0.050	0.048	0.051	0.050	n/a
		1-ß	0.999	*					
	B	α	0.050	0.050	0.049	0.048	0.050	0.048	n/a
		1-ß	n/a	n/a	n/a	0.999	*		
	H-S	α	0.036	0.021	0.021	0.020	0.034	0.014	n/a
		1-ß	0.999	*					
	B-S	α	0.049	0.050	0.051	0.049	0.049	0.050	n/a
		1-ß	0.999	*					

*power > .9999

α=.05

Table 16

Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with 3 degrees of freedom and n=7.

			Number of nonnull effects						
c	Stat	Outcome	1	2	3	4	5	6	7
0.20	F	α	0.046	0.044	0.046	0.046	0.048	0.046	n/a
		1-ß	0.133	0.133	0.133	0.136	0.134	0.136	0.135
	B	α	0.047	0.045	0.049	0.048	0.046	0.050	n/a
		1-ß	n/a	n/a	n/a	0.105	0.106	0.109	0.108
	H-S	α	0.047	0.044	0.043	0.042	0.041	0.038	n/a
		1-ß	0.172	0.169	0.162	0.159	0.149	0.143	0.135
	B-S	α	0.055	0.049	0.050	0.050	0.052	0.051	n/a
		1-ß	0.174	0.163	0.173	0.170	0.169	0.176	0.172
0.80	F	α	0.045	0.046	0.047	0.047	0.044	0.048	n/a
		1-ß	0.870	0.870	0.871	0.870	0.871	0.869	0.870
	B	α	0.047	0.047	0.047	0.046	0.047	0.044	n/a
		1-ß	n/a	n/a	n/a	0.749	0.754	0.751	0.753
	H-S	α	0.025	0.013	0.008	0.009	0.021	0.131	n/a
		1-ß	0.972	0.937	0.890	0.819	0.690	0.517	0.316
	B-S	α	0.050	0.050	0.050	0.050	0.048	0.052	n/a
		1-ß	0.972	0.970	0.969	0.967	0.959	0.929	0.943

$\alpha=.05$

Table 17

Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom and n=35.

			Number of nonnull effects						
c	Stat	Outcome	1	2	3	4	5	6	7
0.20	F	α	0.047	0.049	0.049	0.048	0.049	0.046	n/a
		1-ß	0.433	0.431	0.432	0.434	0.433	0.435	0.432
	B	α	0.049	0.049	0.049	0.048	0.049	0.049	n/a
		1-ß	n/a	n/a	n/a	0.450	0.445	0.453	0.446
	H-S	α	0.047	0.046	0.043	0.041	0.040	0.041	n/a
		1-ß	0.643	0.629	0.619	0.611	0.591	0.569	0.533
	B-S	α	0.049	0.051	0.050	0.049	0.049	0.049	n/a
		1-ß	0.643	0.632	0.641	0.643	0.633	0.637	0.632
0.80	F	α	0.049	0.049	0.048	0.048	0.046	0.046	n/a
		1-ß	0.996	0.999	0.998	0.998	0.998	0.999	0.999
	B	α	0.050	0.051	0.048	0.049	0.046	0.049	n/a
		1-ß	n/a	n/a	n/a	0.999	*		
	H-S	α	0.025	0.013	0.015	0.033	0.191	0.869	n/a
		1-ß	0.999	0.999	0.999	0.999	0.999	0.999	0.970
	B-S	α	0.050	0.051	0.050	0.050	0.044	0.050	n/a
		1-ß	0.999	*					

*power > .9999

α=.05

Table 18
Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution and n=7.

c	Stat	Outcome	Number of nonnull effects 1	2	3	4	5	6	7
0.20	F	α	0.047	0.047	0.047	0.048	0.048	0.049	n/a
		1-ß	0.119	0.118	0.120	0.121	0.122	0.120	0.122
	B	α	0.046	0.047	0.044	0.045	0.049	0.047	n/a
		1-ß	n/a	n/a	n/a	0.095	0.095	0.094	0.095
	H-S	α	0.046	0.045	0.042	0.042	0.040	0.048	n/a
		1-ß	0.124	0.196	0.183	0.171	0.159	0.146	0.130
	B-S	α	0.049	0.051	0.052	0.052	0.051	0.052	n/a
		1-ß	0.216	0.200	0.188	0.201	0.198	0.192	0.165
0.80	F	α	0.048	0.049	0.048	0.049	0.047	0.047	n/a
		1-ß	0.831	0.836	0.834	0.835	0.834	0.835	0.834
	B	α	0.049	0.047	0.048	0.048	0.045	0.043	n/a
		1-ß	n/a	n/a	n/a	0.694	0.688	0.690	0.687
	H-S	α	0.026	0.016	0.011	0.011	0.024	0.147	n/a
		1-ß	0.959	0.914	0.855	0.779	0.650	0.495	0.308
	B-S	α	0.051	0.052	0.052	0.052	0.047	0.047	n/a
		1-ß	0.960	0.964	0.956	0.960	0.943	0.899	0.913

α=.05

Table 19

Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution and n=35.

c	Stat	Outcome	Number of nonnull effects 1	2	3	4	5	6	7
0.20	F	α	0.050	0.049	0.049	0.050	0.048	0.049	n/a
		1-ß	0.386	0.392	0.389	0.391	0.392	0.393	0.392
	B	α	0.050	0.050	0.049	0.049	0.048	0.050	n/a
		1-ß	n/a	n/a	n/a	0.387	0.383	0.387	0.384
	H-S	α	0.048	0.044	0.044	0.043	0.042	0.052	n/a
		1-ß	0.751	0.713	0.687	0.657	0.621	0.577	0.513
	B-S	α	0.051	0.049	0.051	0.050	0.049	0.048	n/a
		1-ß	0.751	0.738	0.731	0.735	0.715	0.681	0.671
0.80	F	α	0.050	0.050	0.051	0.051	0.051	0.051	n/a
		1-ß	0.999	*					
	B	α	0.049	0.049	0.050	0.050	0.049	0.051	n/a
		1-ß	n/a	n/a	n/a	0.999	*		
	H-S	α	0.026	0.016	0.009	0.040	0.198	0.837	n/a
		1-ß	0.999	0.999	0.999	0.999	0.999	0.999	0.853
	B-S	α	0.050	0.050	0.051	0.052	0.048	0.045	n/a
		1-ß	0.999	*					

*power > .9999

α=.05

Table 20

Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set and n=7.

c	Stat	Outcome	Number of nonnull effects 1	2	3	4	5	6	7
0.20	F	α	0.050	0.050	0.050	0.051	0.050	0.051	n/a
		1-ß	0.109	0.111	0.114	0.111	0.112	0.111	0.112
	B	α	0.044	0.046	0.046	0.046	0.046	0.048	n/a
		1-ß	n/a	n/a	n/a	0.088	0.088	0.084	0.087
	H-S	α	0.048	0.047	0.046	0.045	0.044	0.044	n/a
		1-ß	0.132	0.120	0.117	0.111	0.110	0.103	0.096
	B-S	α	0.050	0.050	0.051	0.050	0.050	0.051	n/a
		1-ß	0.132	0.126	0.124	0.119	0.119	0.118	0.115
0.80	F	α	0.049	0.051	0.052	0.050	0.050	0.049	n/a
		1-ß	0.836	0.835	0.836	0.835	0.836	0.837	0.835
	B	α	0.043	0.045	0.047	0.047	0.047	0.042	n/a
		1-ß	n/a	n/a	n/a	0.665	0.670	0.670	0.668
	H-S	α	0.034	0.026	0.019	0.015	0.014	0.025	n/a
		1-ß	0.813	0.744	0.694	0.641	0.556	0.454	0.329
	B-S	α	0.049	0.051	0.040	0.049	0.050	0.049	n/a
		1-ß	0.813	0.809	0.819	0.805	0.790	0.758	0.772

$\alpha=.05$

Table 21

Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi modal lumpy data set and n=35.

c	Stat	Outcome	Number of nonnull effects 1	2	3	4	5	6	7
0.20	F	α	0.050	0.050	0.052	0.050	0.049	0.049	n/a
		1-ß	0.383	0.384	0.384	0.383	0.382	0.385	0.383
	B	α	0.049	0.049	0.050	0.049	0.049	0.049	n/a
		1-ß	n/a	n/a	n/a	0.360	0.360	0.363	0.361
	H-S	α	0.049	0.047	0.046	0.045	0.045	0.045	n/a
		1-ß	0.479	0.428	0.413	0.406	0.391	0.375	0.332
	B-S	α	0.050	0.049	0.049	0.050	0.050	0.050	n/a
		1-ß	0.479	0.459	0.449	0.435	0.427	0.419	0.410
0.80	F	α	0.052	0.051	0.050	0.049	0.050	0.048	n/a
		1-ß	0.999	*					
	B	α	0.050	0.050	0.049	0.048	0.049	0.048	n/a
		1-ß	n/a	n/a	n/a	0.999	*		
	H-S	α	0.037	0.026	0.017	0.021	0.044	0.120	n/a
		1-ß	0.999	*					
	B-S	α	0.052	0.052	0.050	0.050	0.049	0.046	n/a
		1-ß	0.999	*					

*power > .9999

α=.05

Table 22
Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set and n=7.

c	Stat	Outcome	Number of nonnull effects 1	2	3	4	5	6	7
0.20	F	α	0.049	0.051	0.048	0.050	0.049	0.052	n/a
		1-ß	0.116	0.117	0.117	0.117	0.117	0.115	0.115
	B	α	0.046	0.046	0.045	0.044	0.046	0.045	n/a
		1-ß	n/a	n/a	n/a	0.088	0.088	0.089	0.088
	H-S	α	0.044	0.043	0.040	0.040	0.042	0.056	n/a
		1-ß	0.319	0.252	0.217	0.192	0.167	0.151	0.120
	B-S	α	0.050	0.051	0.051	0.053	0.052	0.057	n/a
		1-ß	0.325	0.245	0.211	0.158	0.161	0.165	0.165
0.80	F	α	0.050	0.048	0.050	0.050	0.049	0.050	n/a
		1-ß	0.835	0.835	0.829	0.832	0.834	0.832	0.833
	B	α	0.045	0.042	0.048	0.046	0.045	0.044	n/a
		1-ß	n/a	n/a	n/a	0.677	0.679	0.678	0.679
	H-S	α	0.028	0.018	0.014	0.014	0.026	0.106	n/a
		1-ß	0.934	0.873	0.801	0.741	0.607	0.473	0.312
	B-S	α	0.051	0.050	0.052	0.053	0.056	0.061	n/a
		1-ß	0.933	0.934	0.931	0.927	0.904	0.836	0.876

α=.05

Table 23
Type I error rates and power levels of the ANOVA F, Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set and n=35.

c	Stat	Outcome	Number of nonnull effects 1	2	3	4	5	6	7
0.20	F	α	0.050	0.050	0.050	0.050	0.049	0.050	n/a
		1-ß	0.388	0.388	0.385	0.384	0.385	0.388	0.387
	B	α	0.051	0.050	0.050	0.049	0.048	0.047	n/a
		1-ß	n/a	n/a	n/a	0.376	0.373	0.377	0.375
	H-S	α	0.046	0.043	0.043	0.045	0.061	0.137	n/a
		1-ß	0.920	0.838	0.774	0.741	0.652	0.583	0.468
	B-S	α	0.051	0.050	0.051	0.051	0.053	0.058	n/a
		1-ß	0.920	0.836	0.783	0.702	0.680	0.652	0.661
0.80	F	α	0.049	0.051	0.050	0.051	0.049	0.052	n/a
		1-ß	0.999	*					
	B	α	0.048	0.050	0.049	0.049	0.048	0.051	n/a
		1-ß	n/a	n/a	n/a	0.999	*		
	H-S	α	0.028	0.019	0.020	0.037	0.153	0.739	n/a
		1-ß	0.999	0.999	0.999	0.999	0.999	0.999	0.968
	B-S	α	0.050	0.052	0.049	0.051	0.055	0.075	n/a
		1-ß	0.999	*					

*power > .9999

α=.05

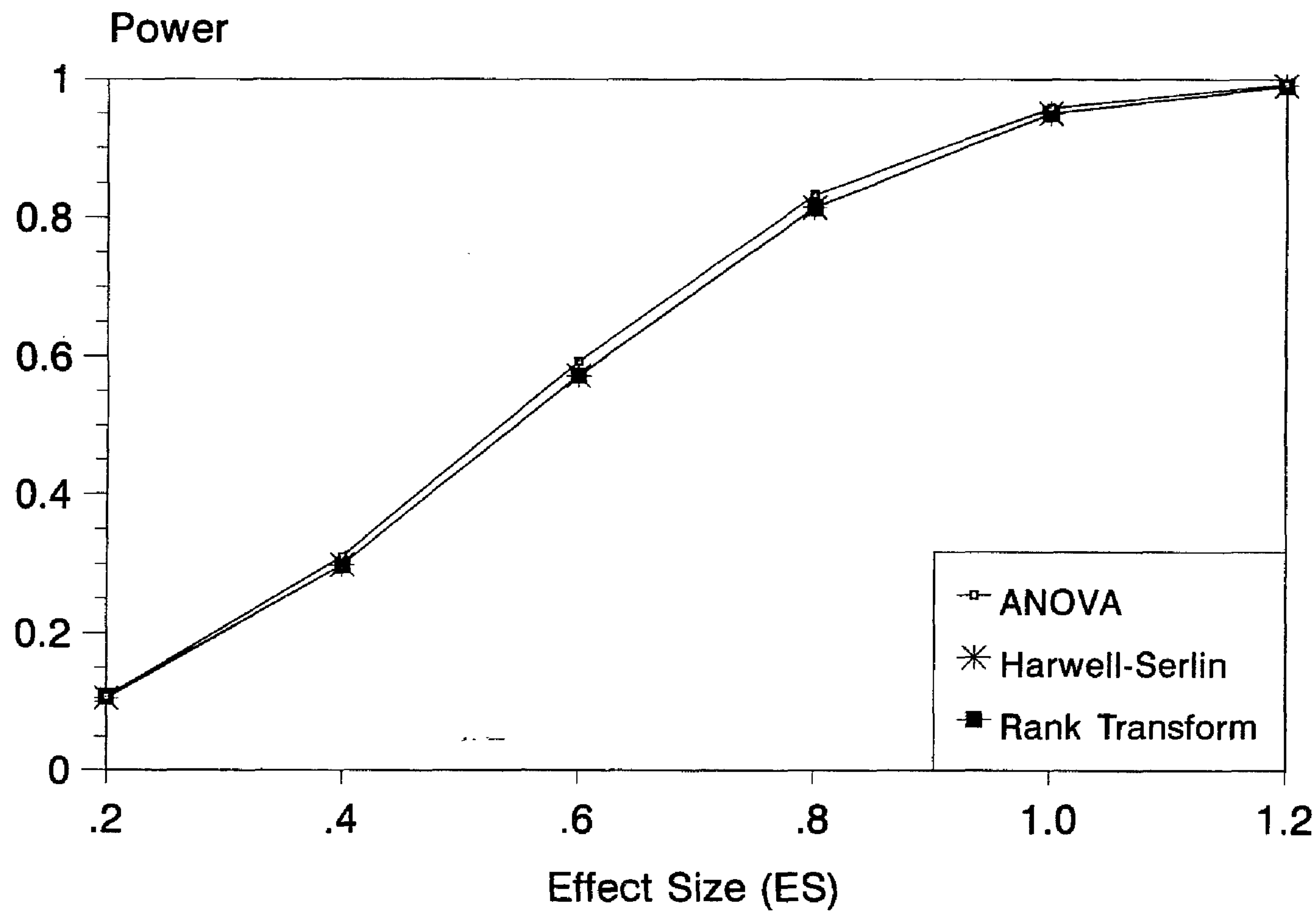

Figure 1. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the Gaussian distribution, alpha=.05 and n=7.

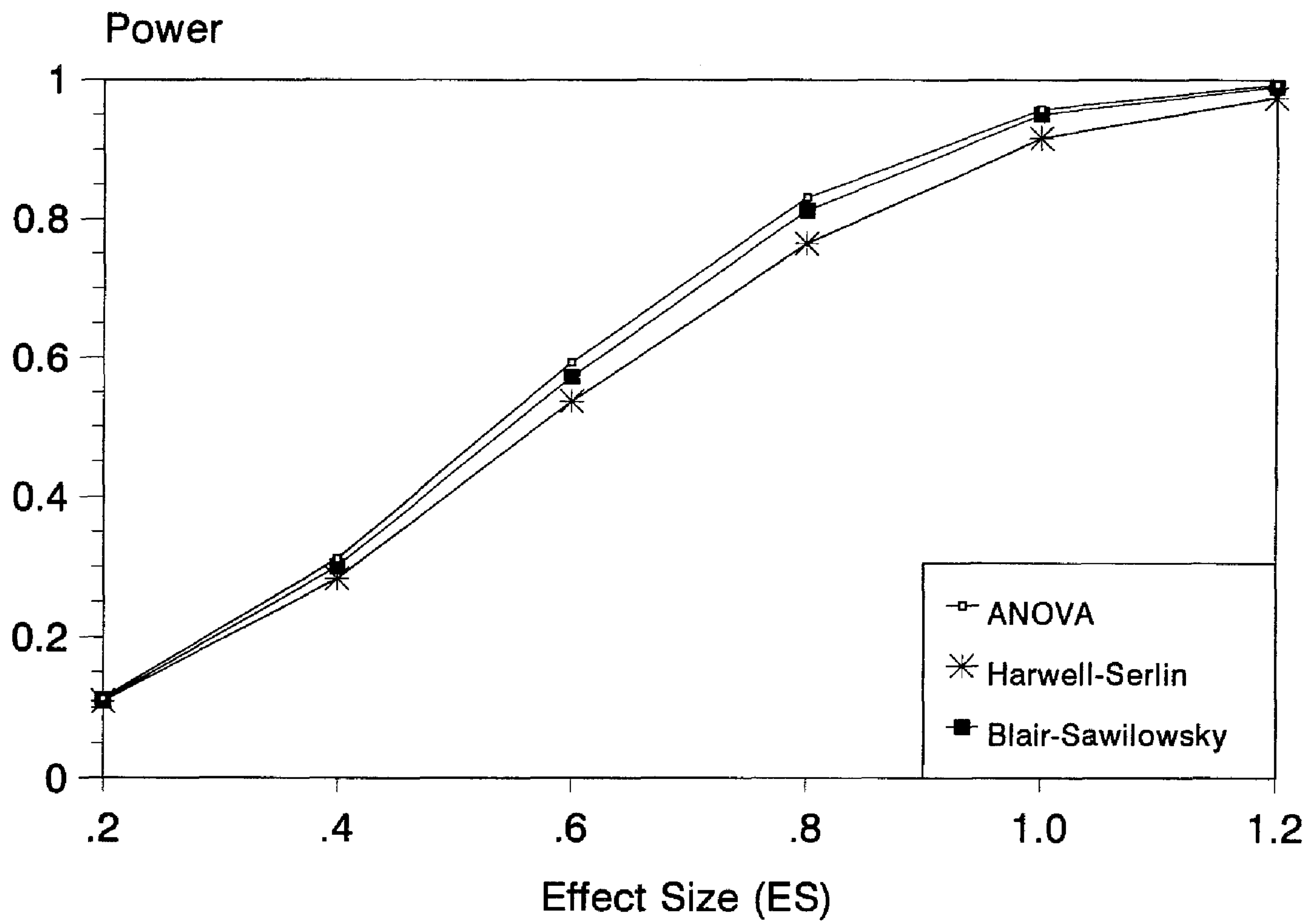

Figure 2. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05, and n=7.

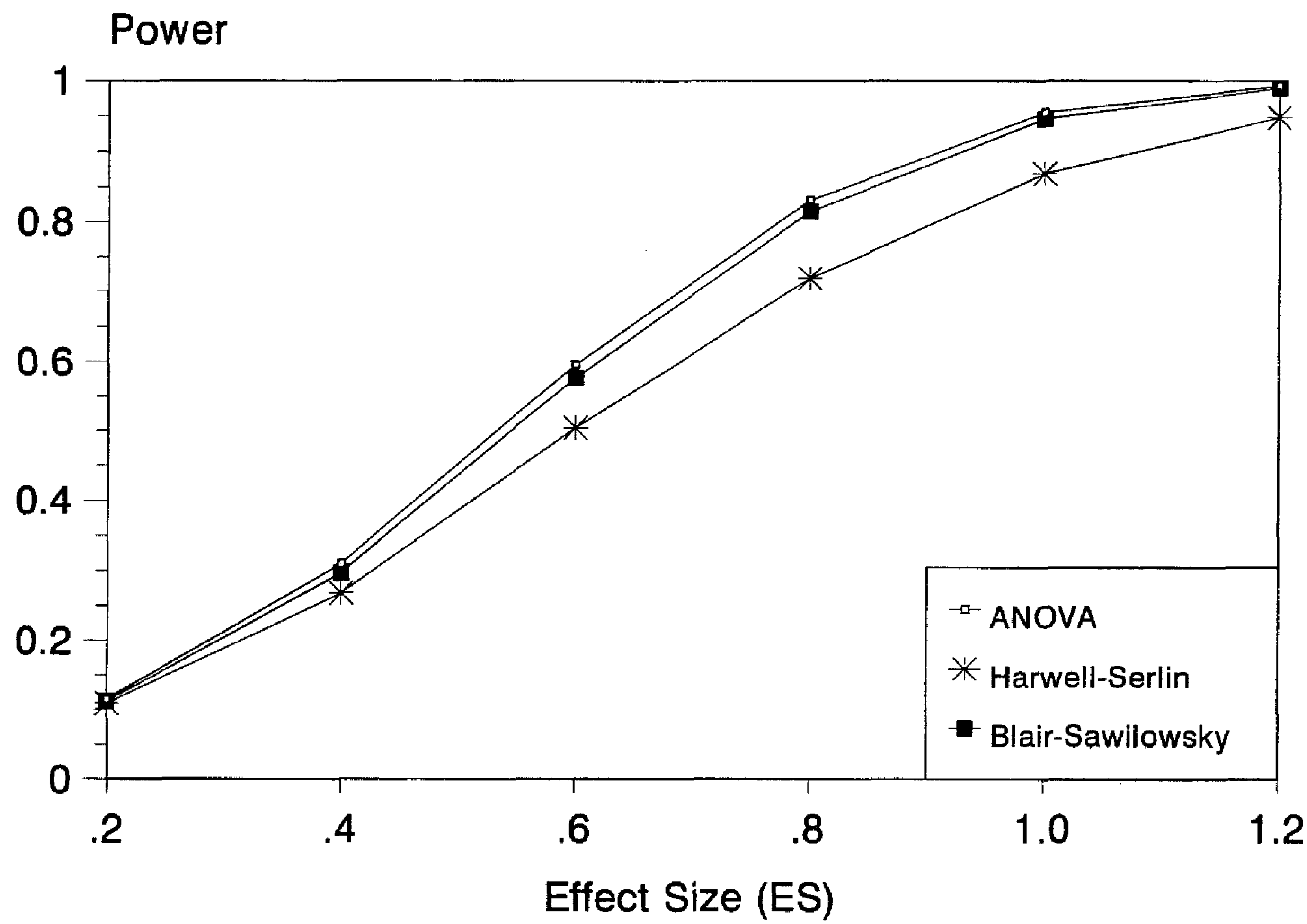

Figure 3. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=7.

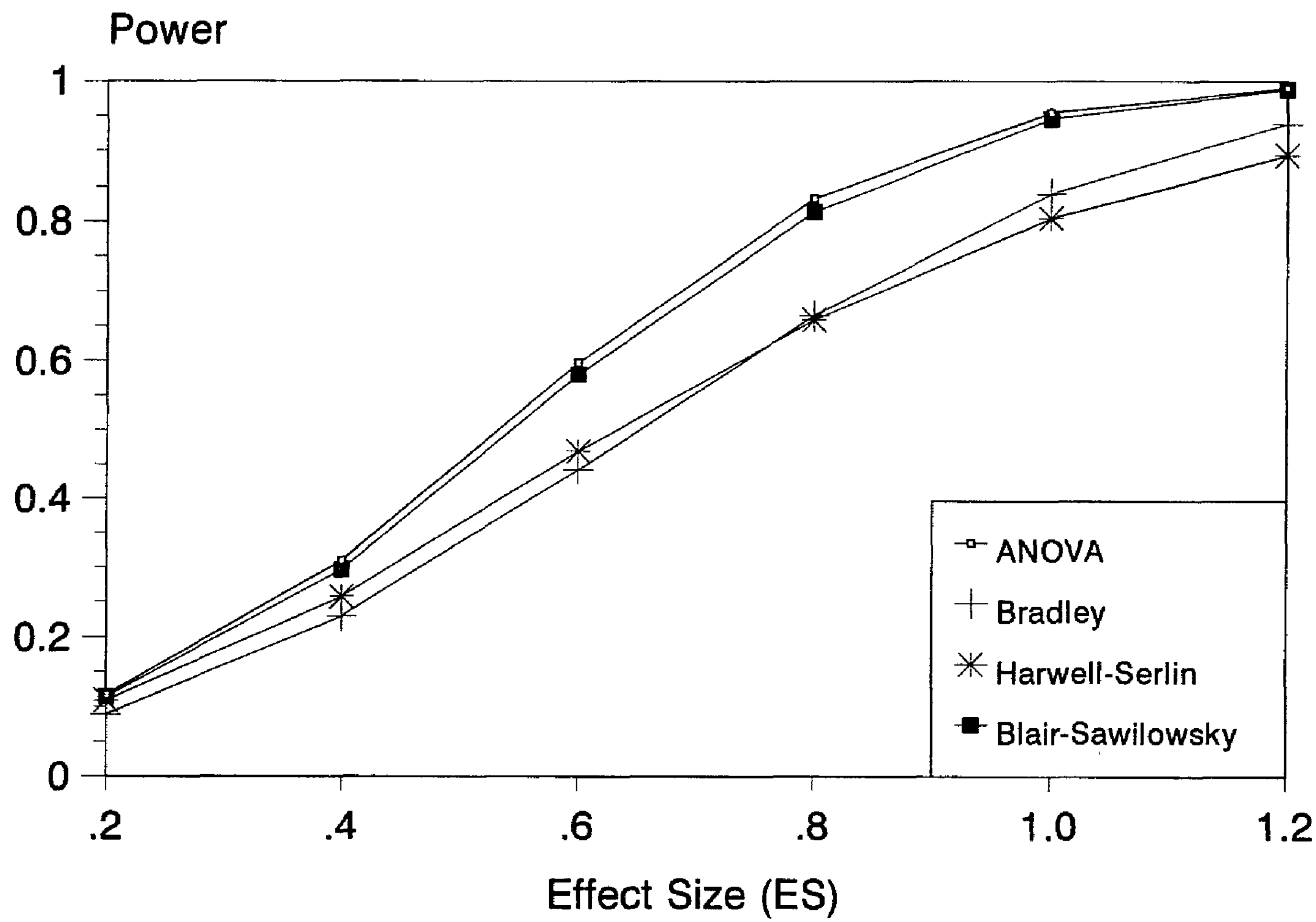

Figure 4. Comparative power of (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=7.

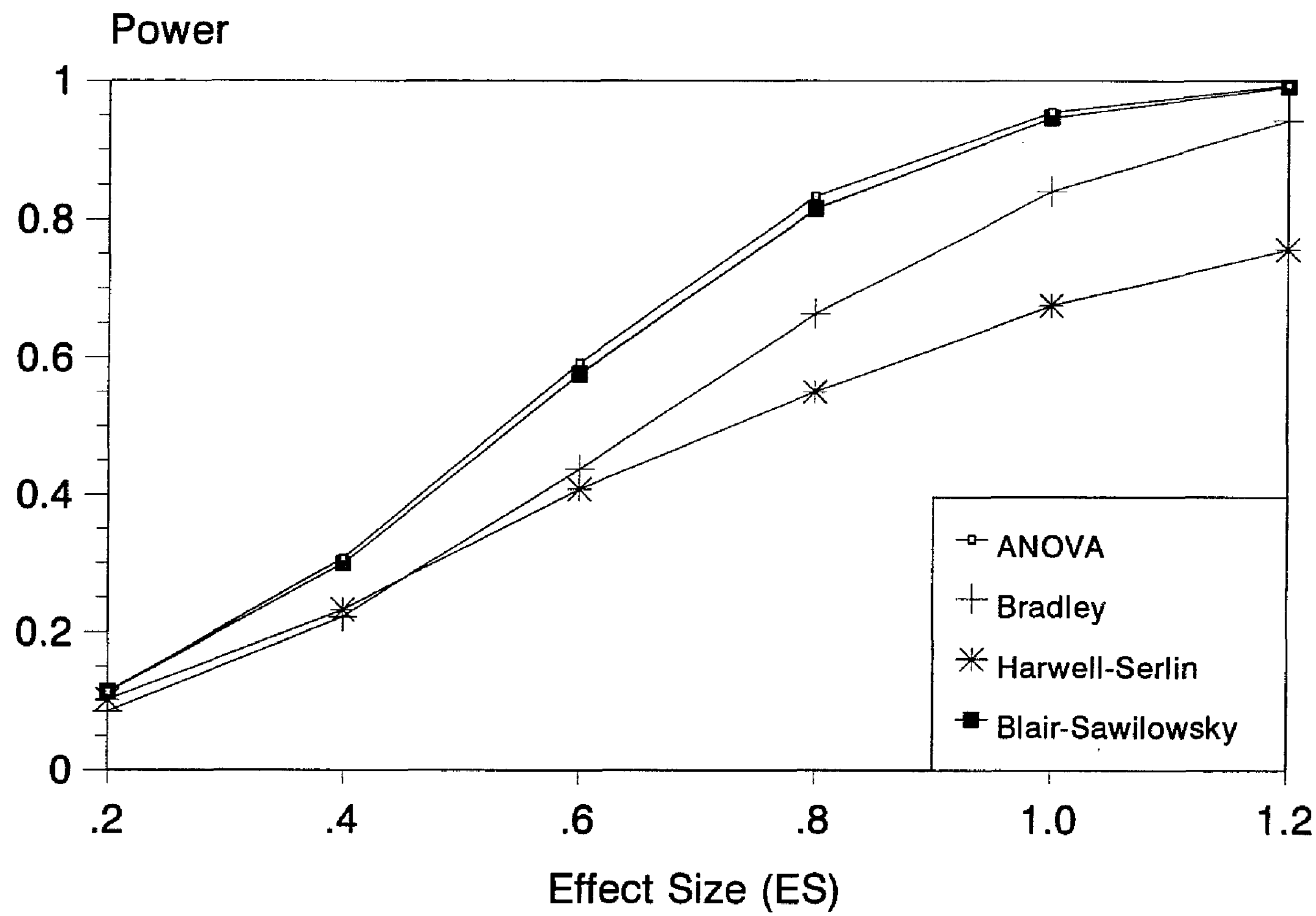

Figure 5. Comparative power of (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=7.

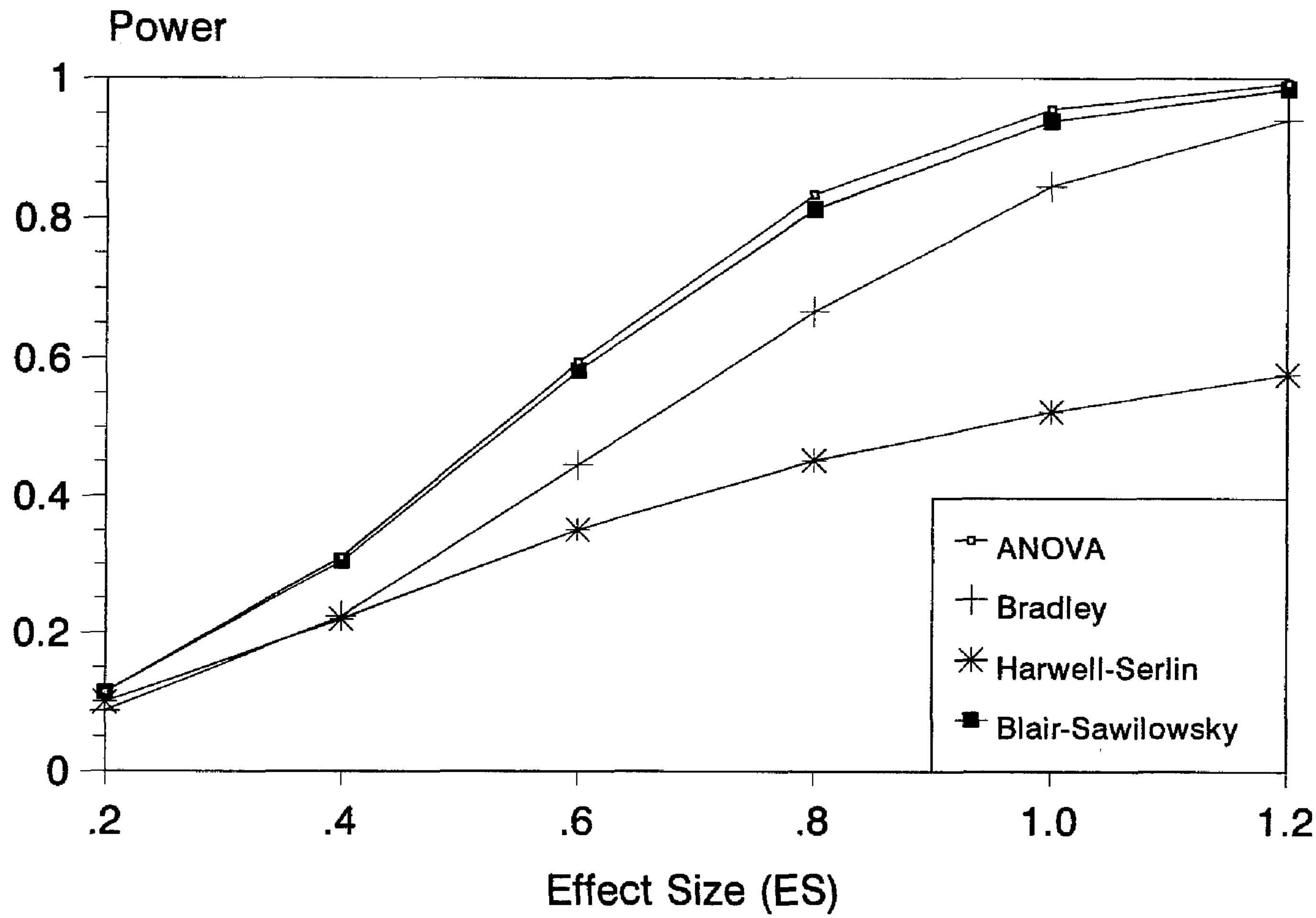

Figure 6. Comparative power of (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=7.

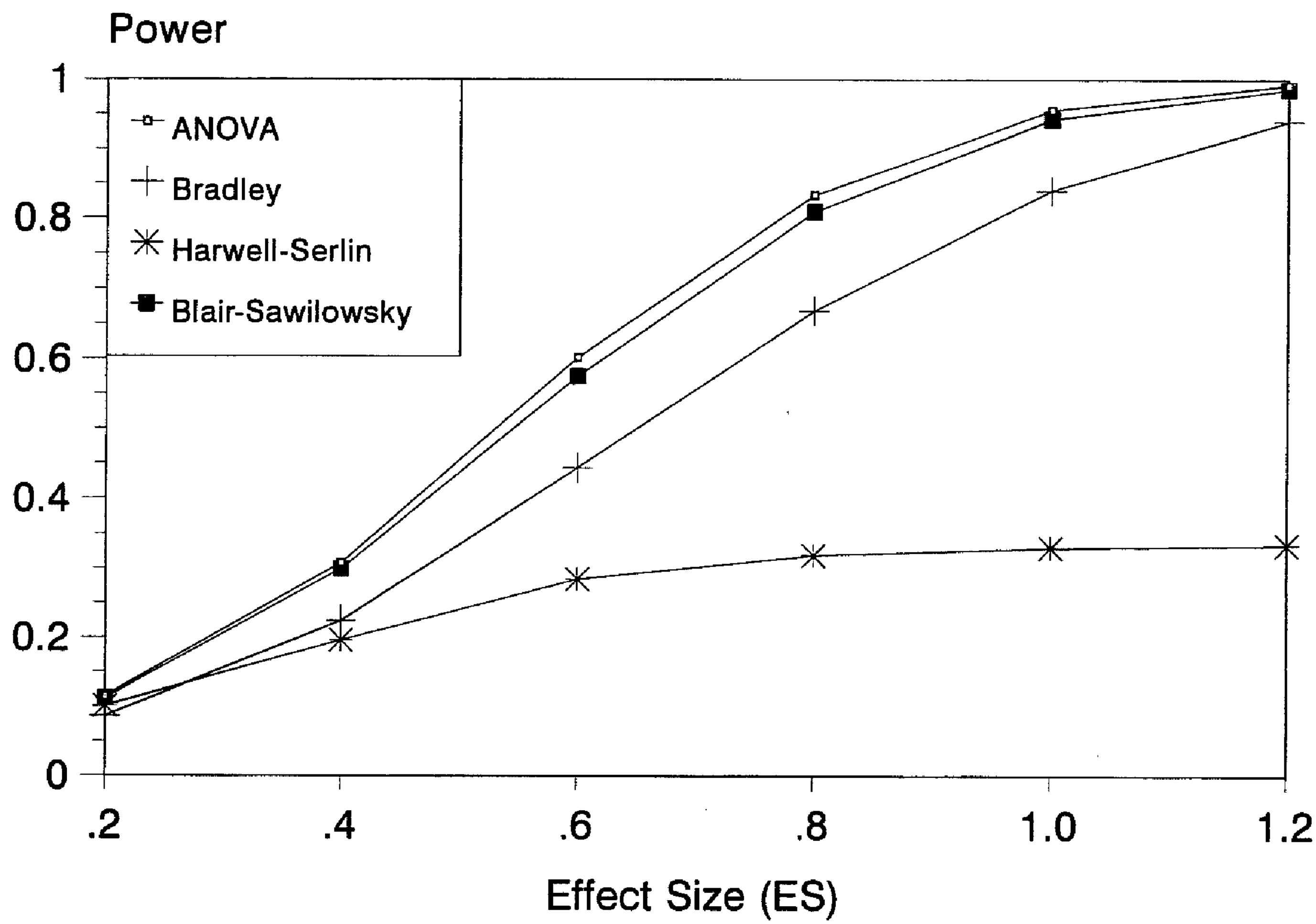

Figure 7. Comparative power of (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=7.

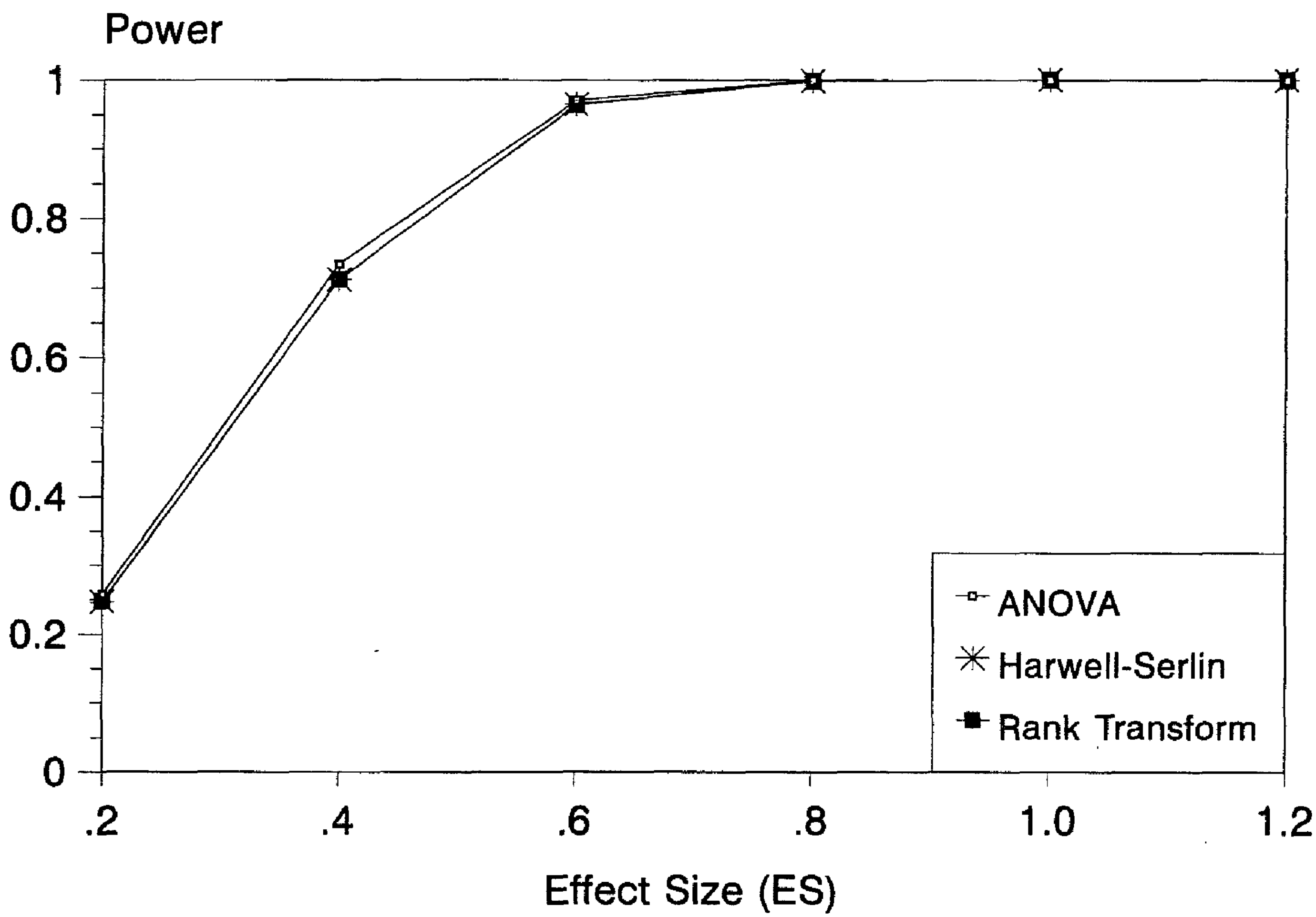

Figure 8. Comparative power of A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the Gaussian distribution, alpha=.05, and n=21.

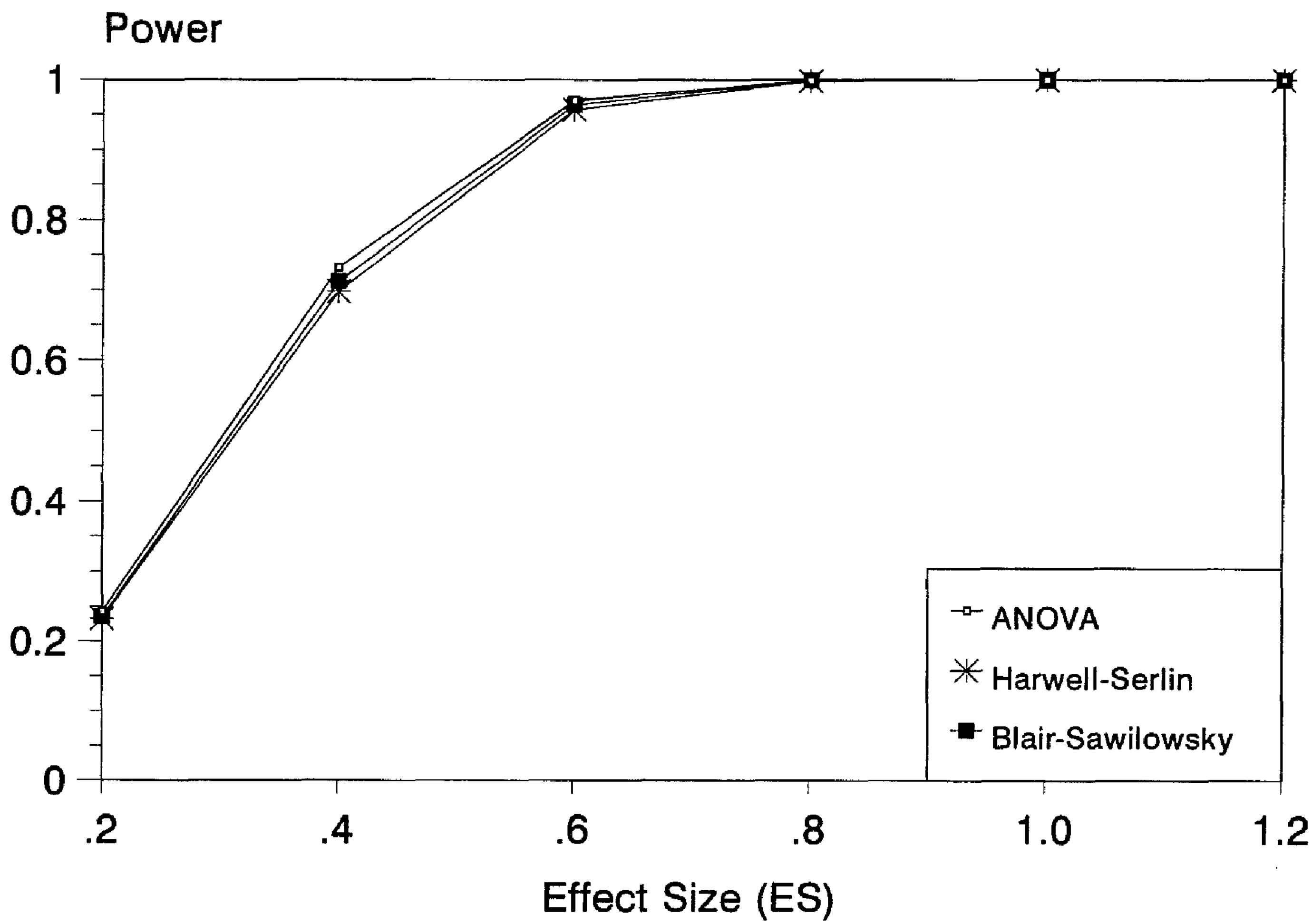

Figure 9. Comparative power of B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=21.

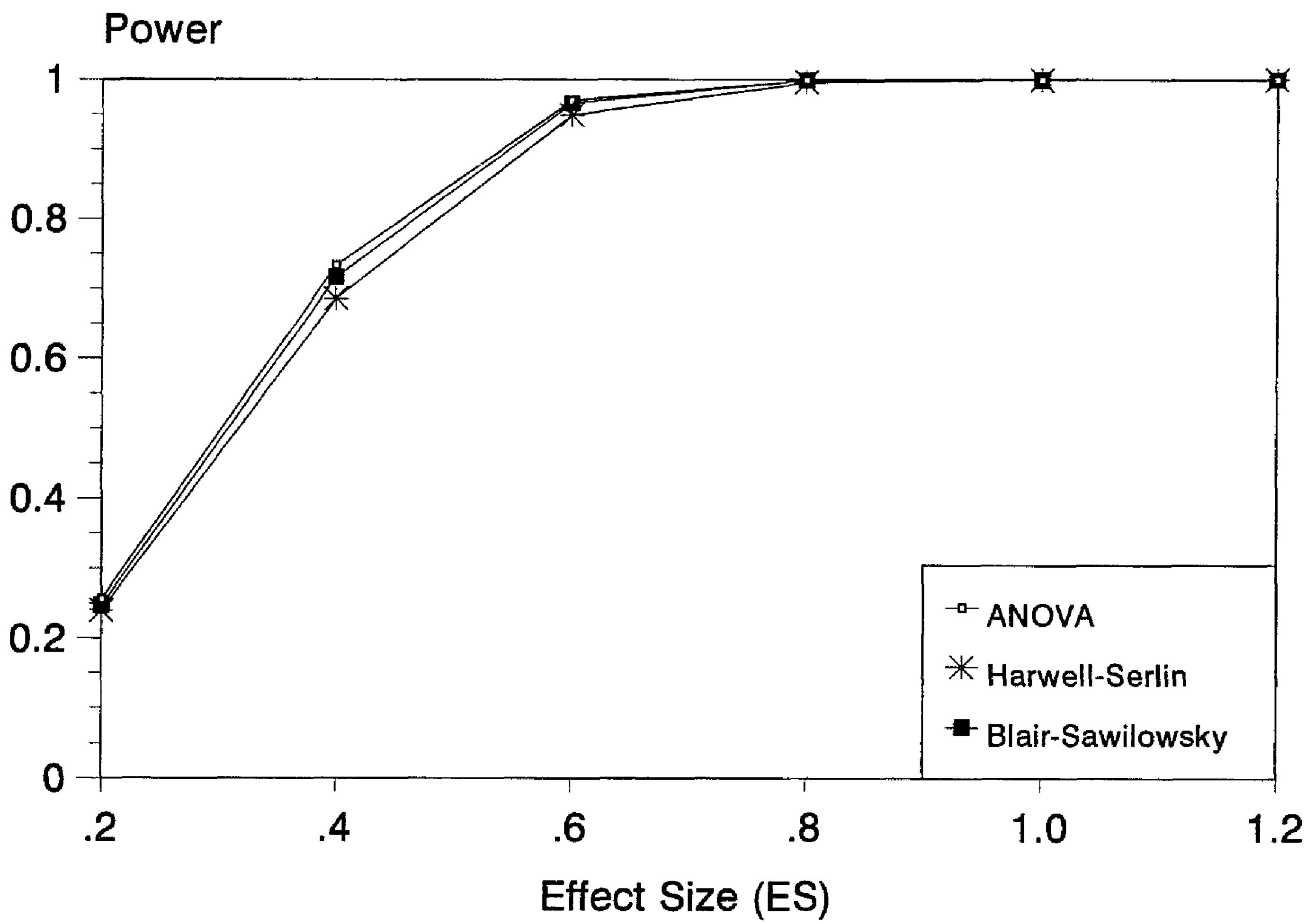

Figure 10. Comparative power of C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=21.

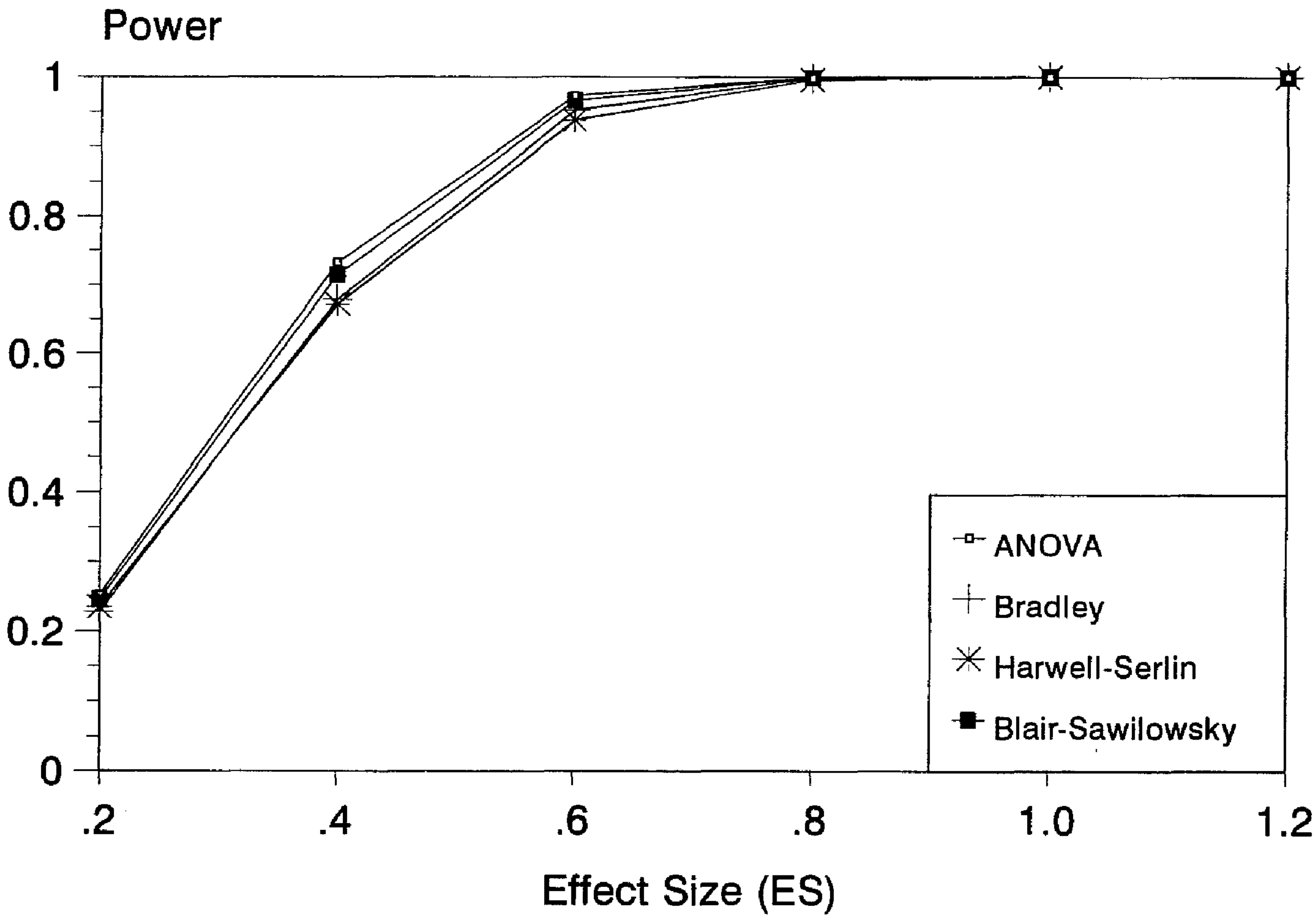

Figure 11. Comparative power of (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=21.

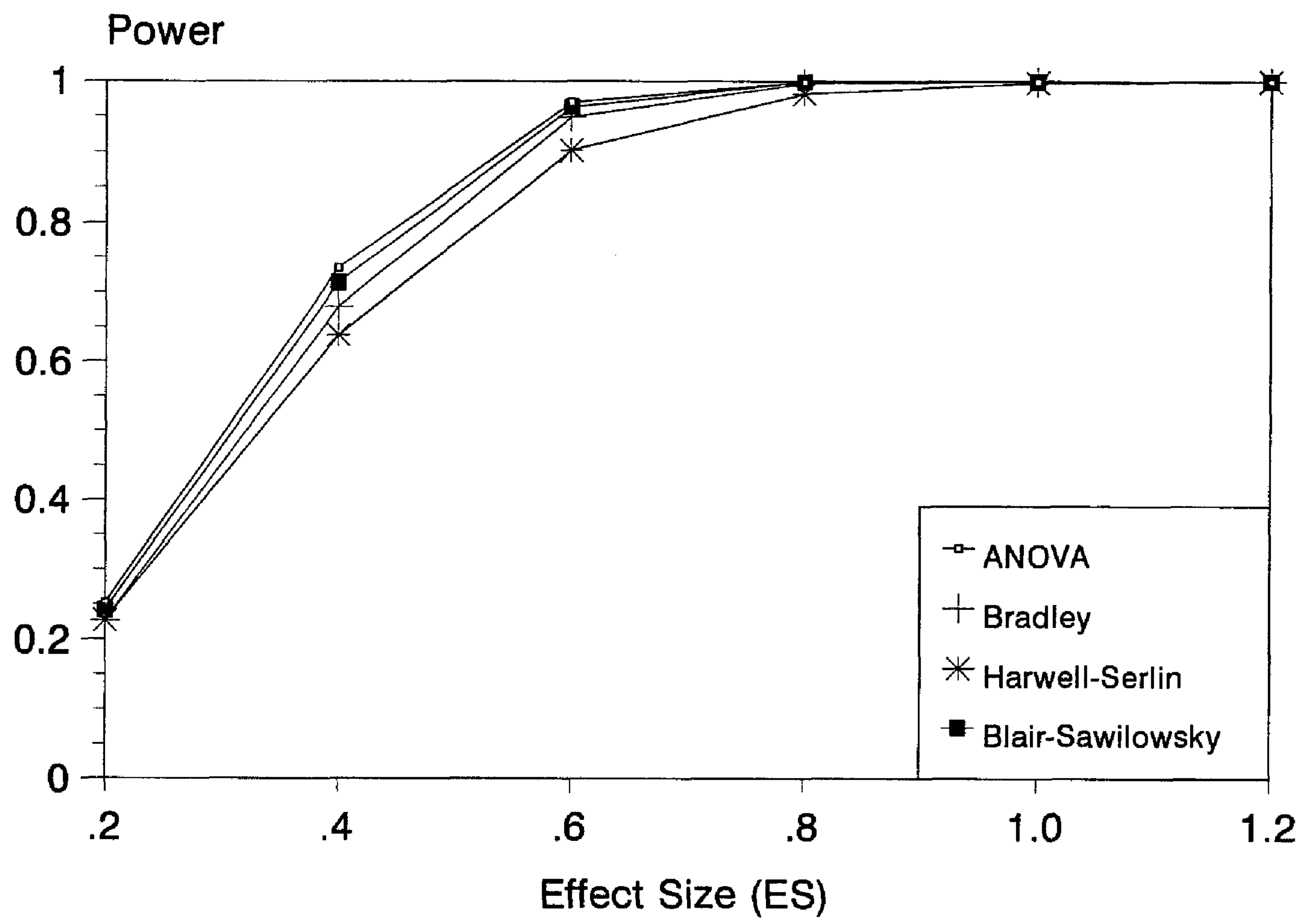

Figure 12. Comparative power of (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=21.

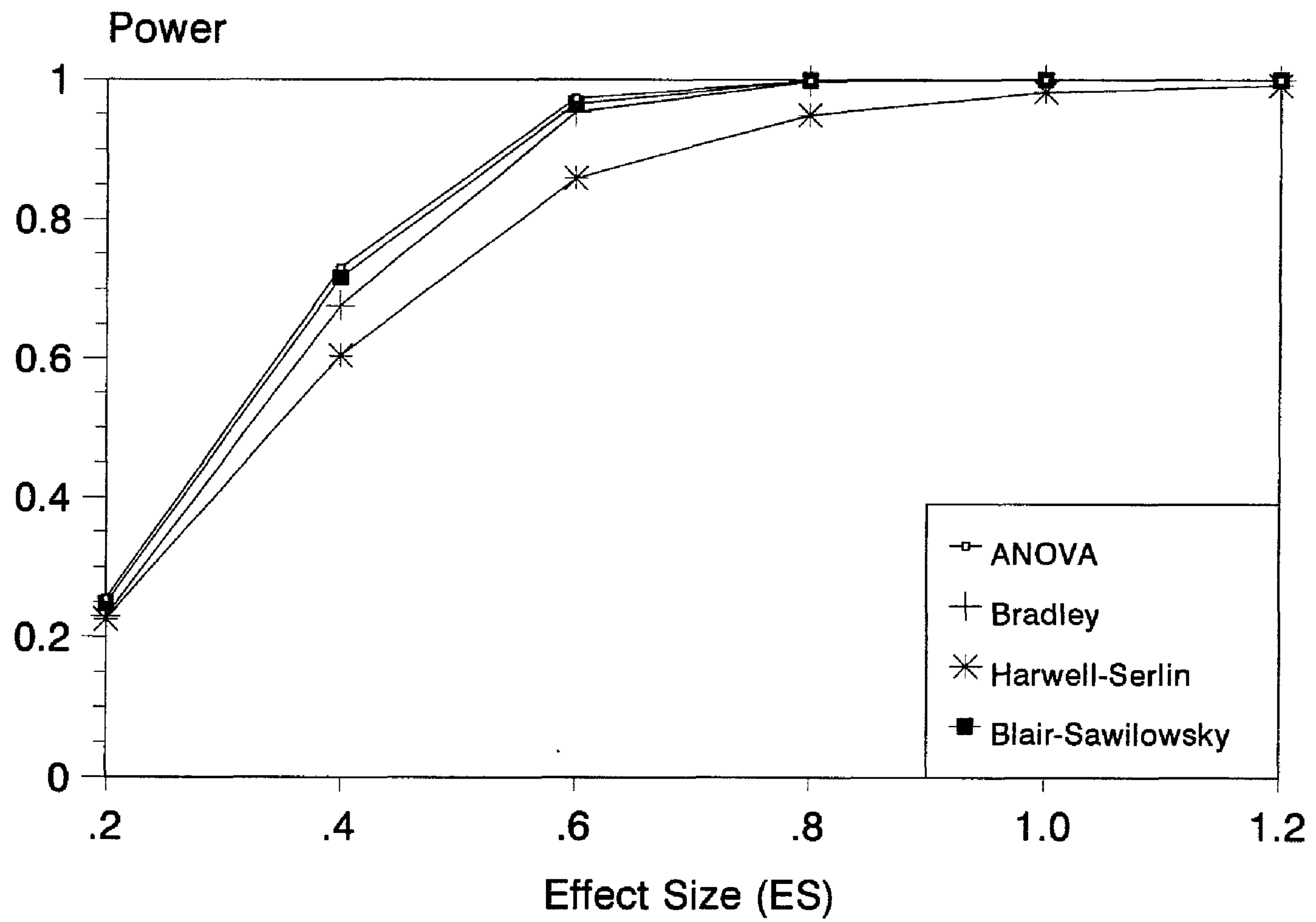

Figure 13. Comparative power of (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05, and n=21.

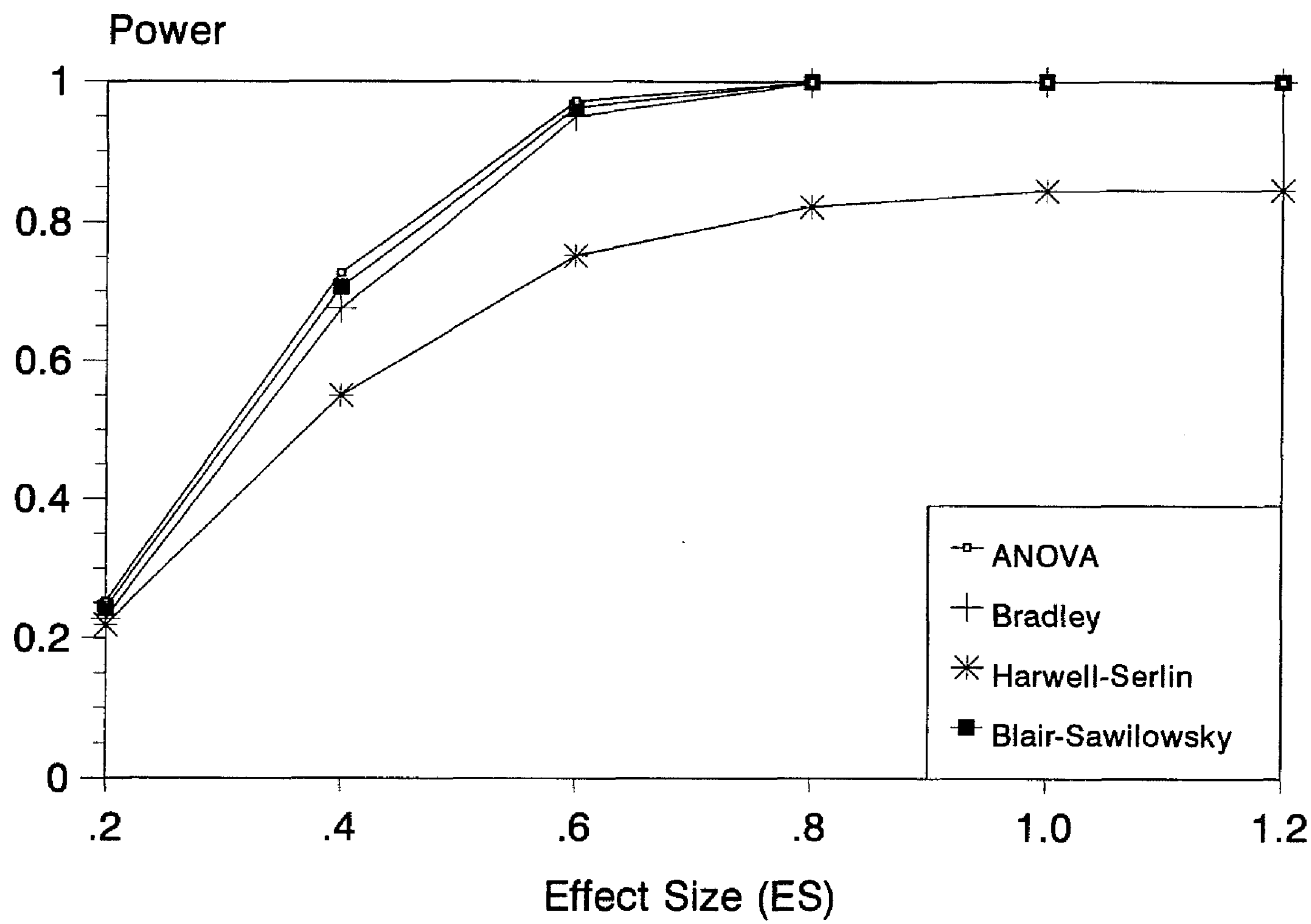

Figure 14. Comparative power of (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=21.

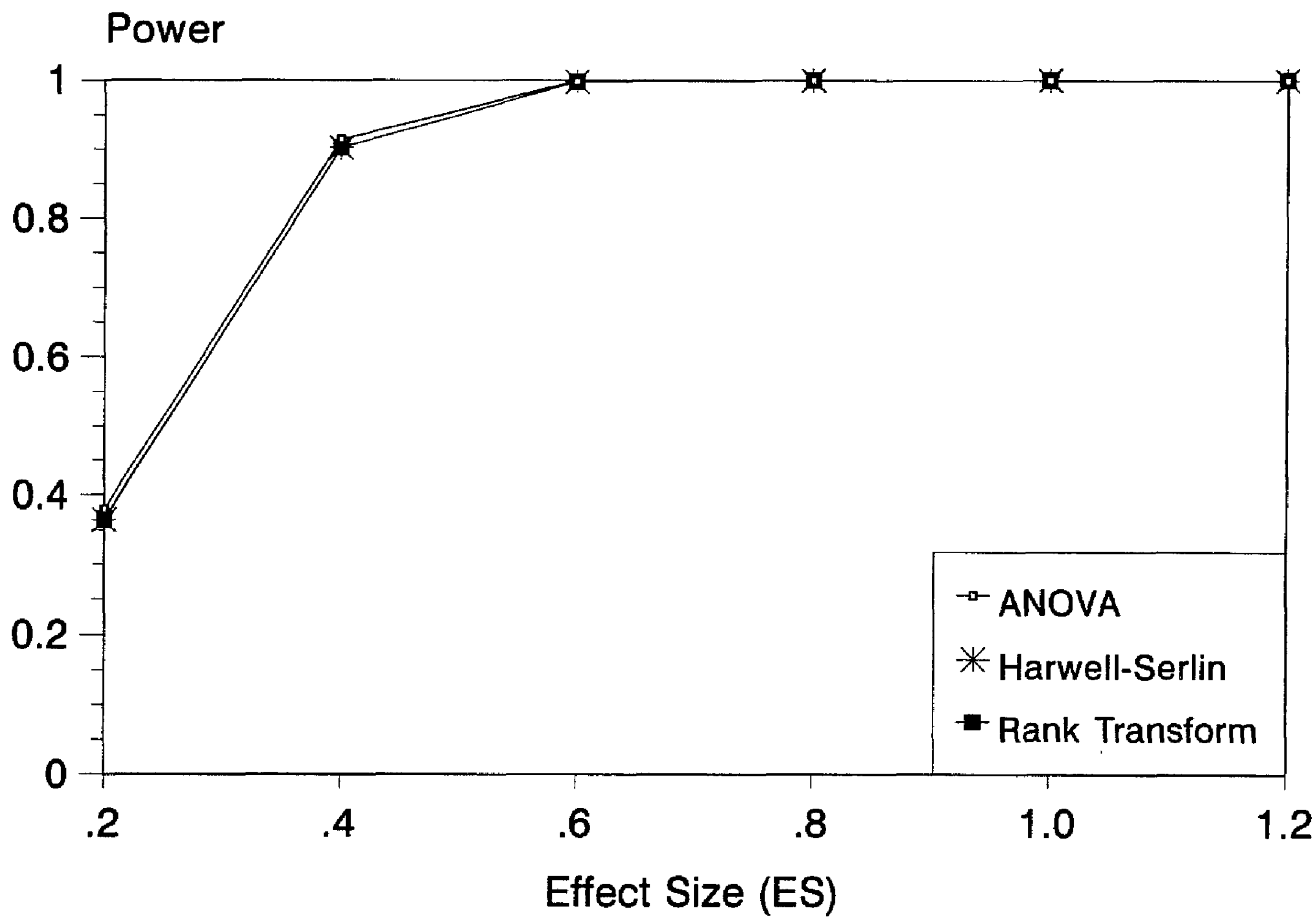

Figure 15. Comparative power of A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the Gaussian distribution, alpha=.05 and n=35.

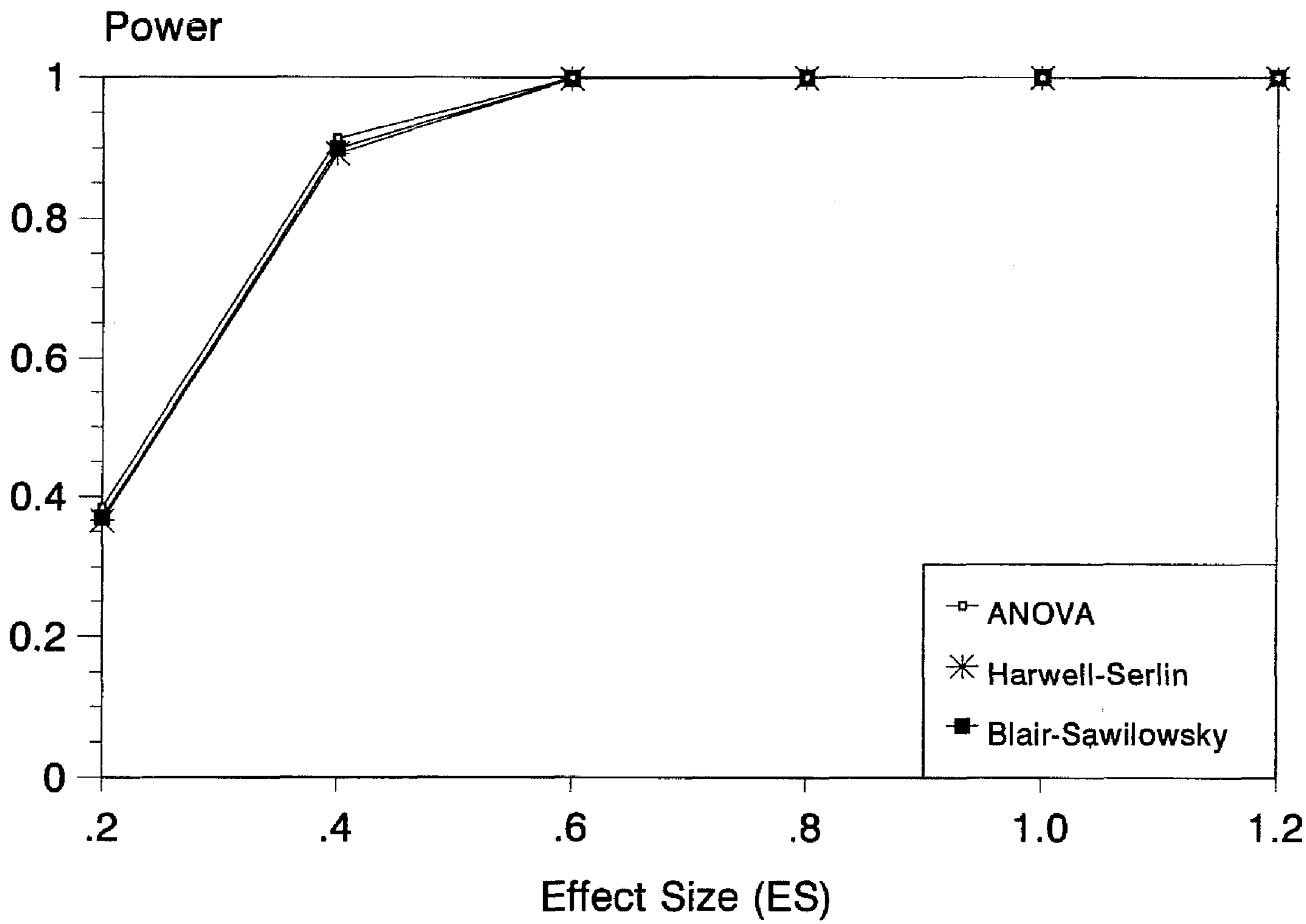

Figure 16. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05, and n=35.

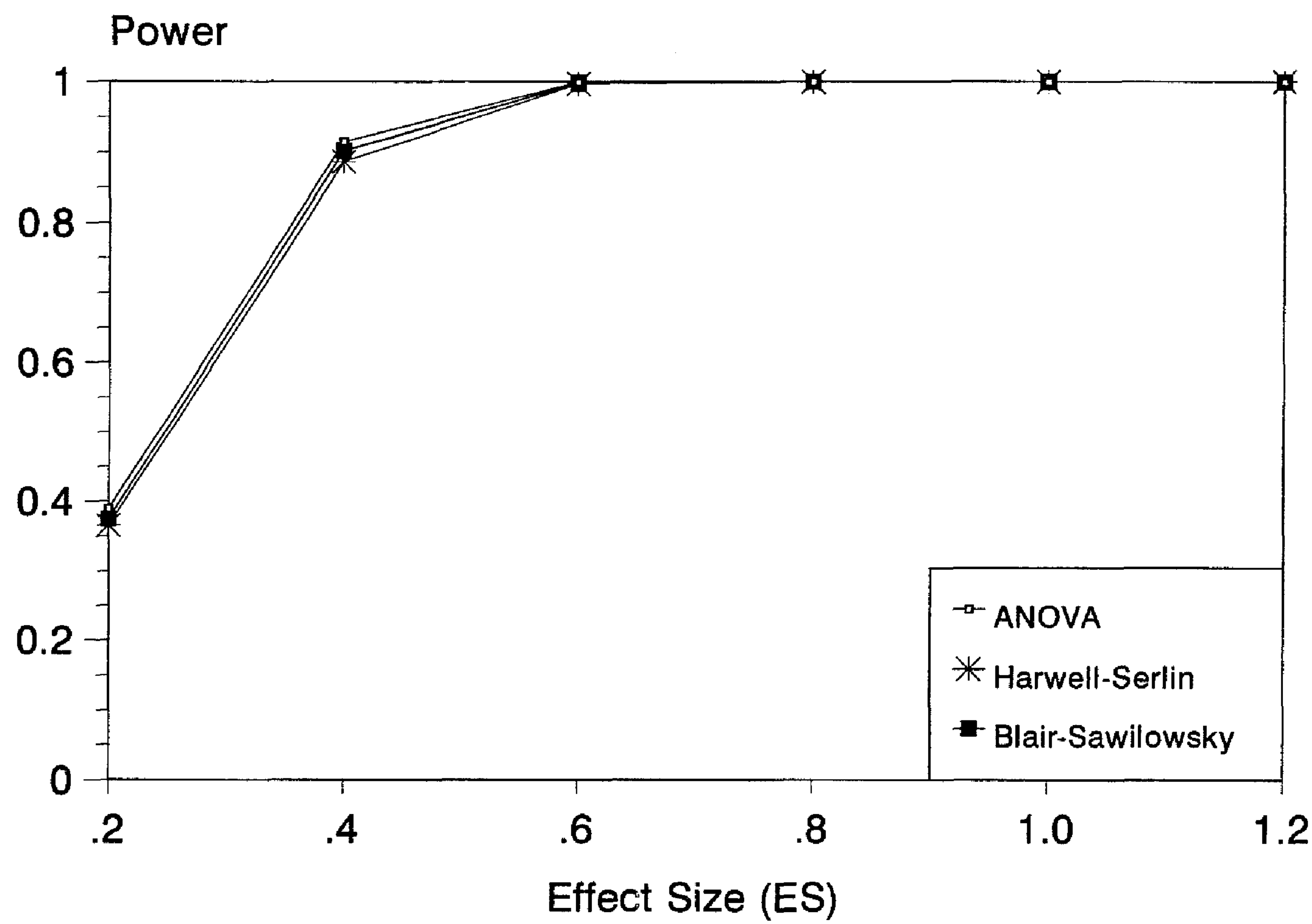

Figure 17. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05, and n=35.

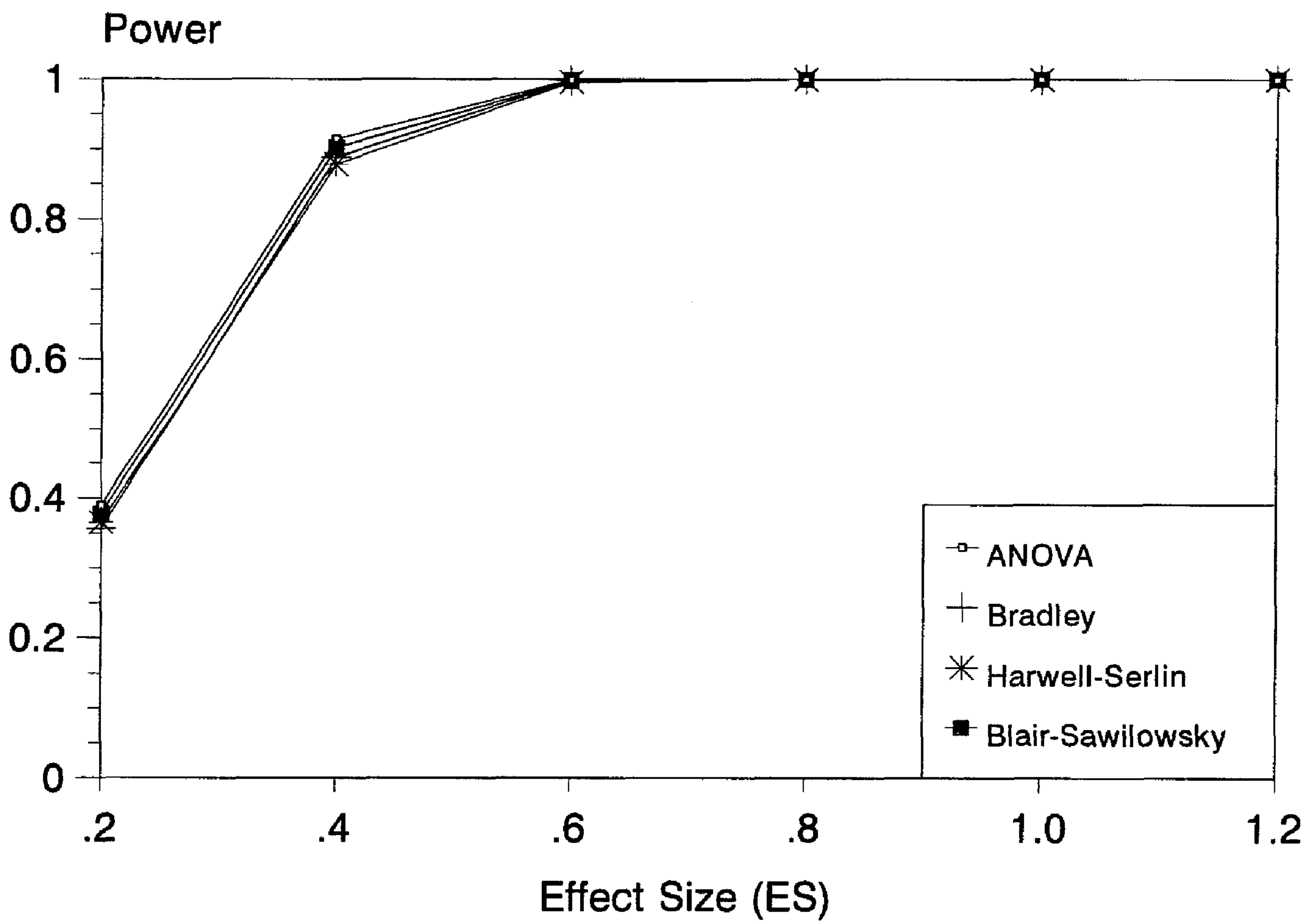

Figure 18. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=35.

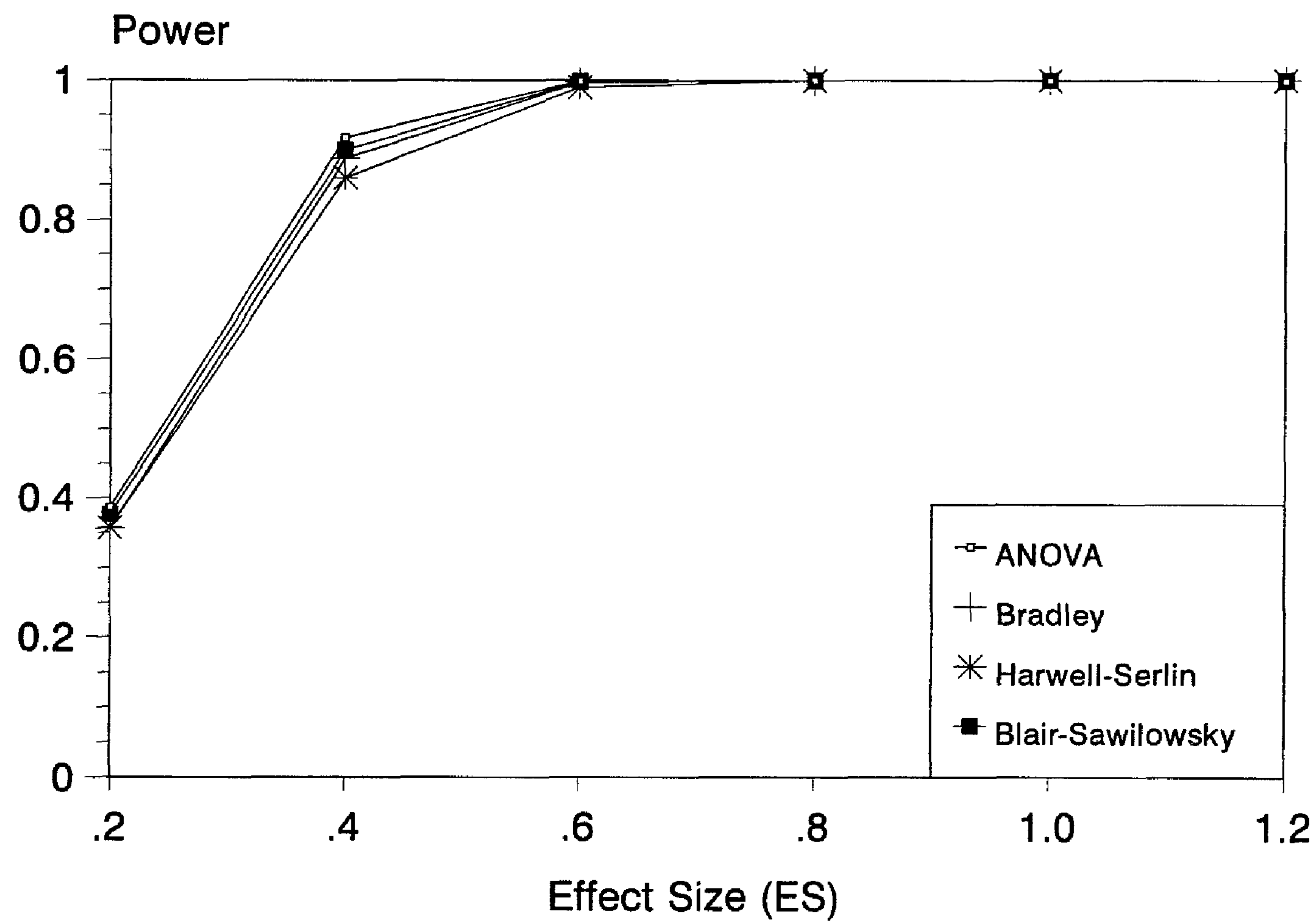

Figure 19. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=35.

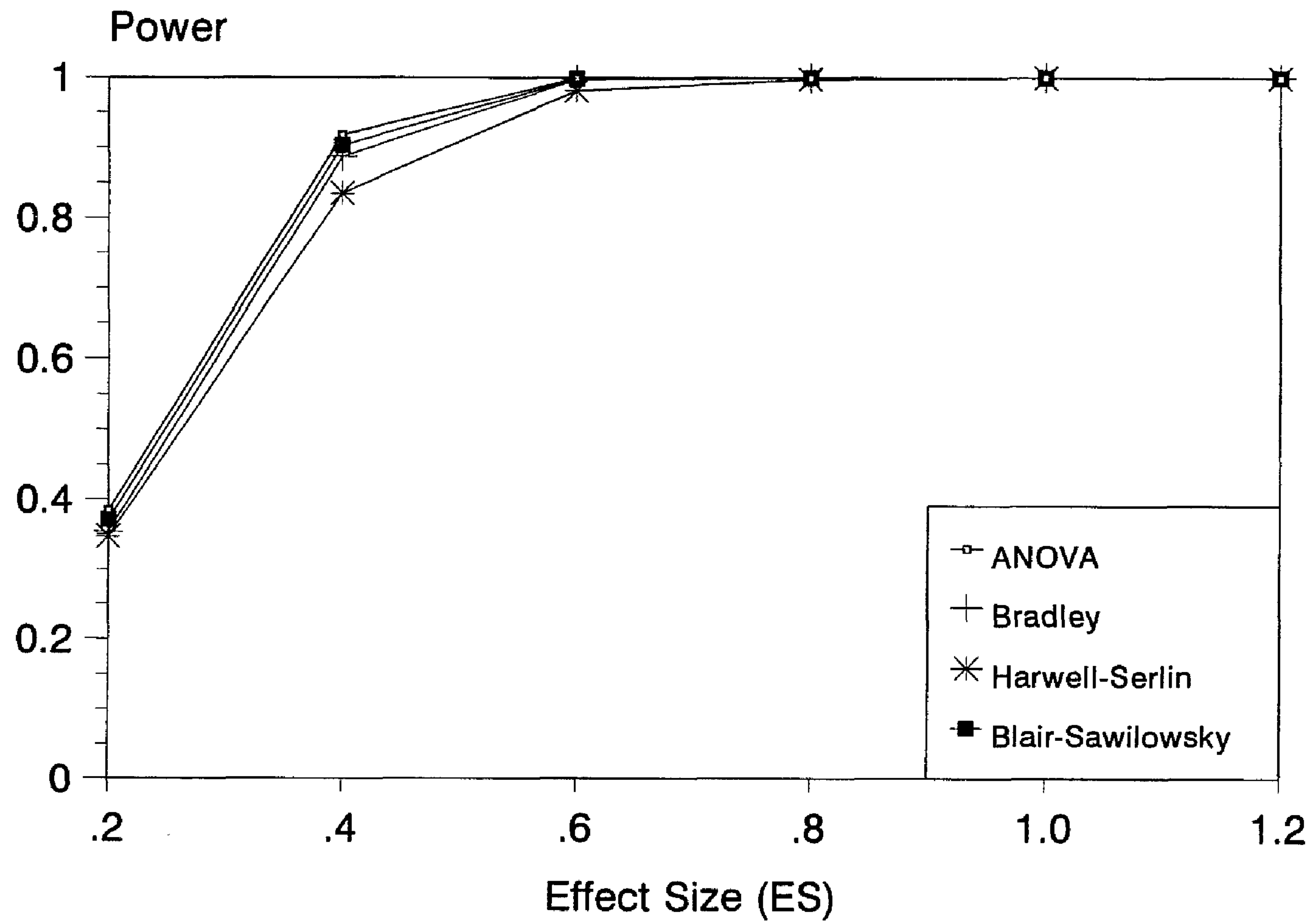

Figure 20. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=35.

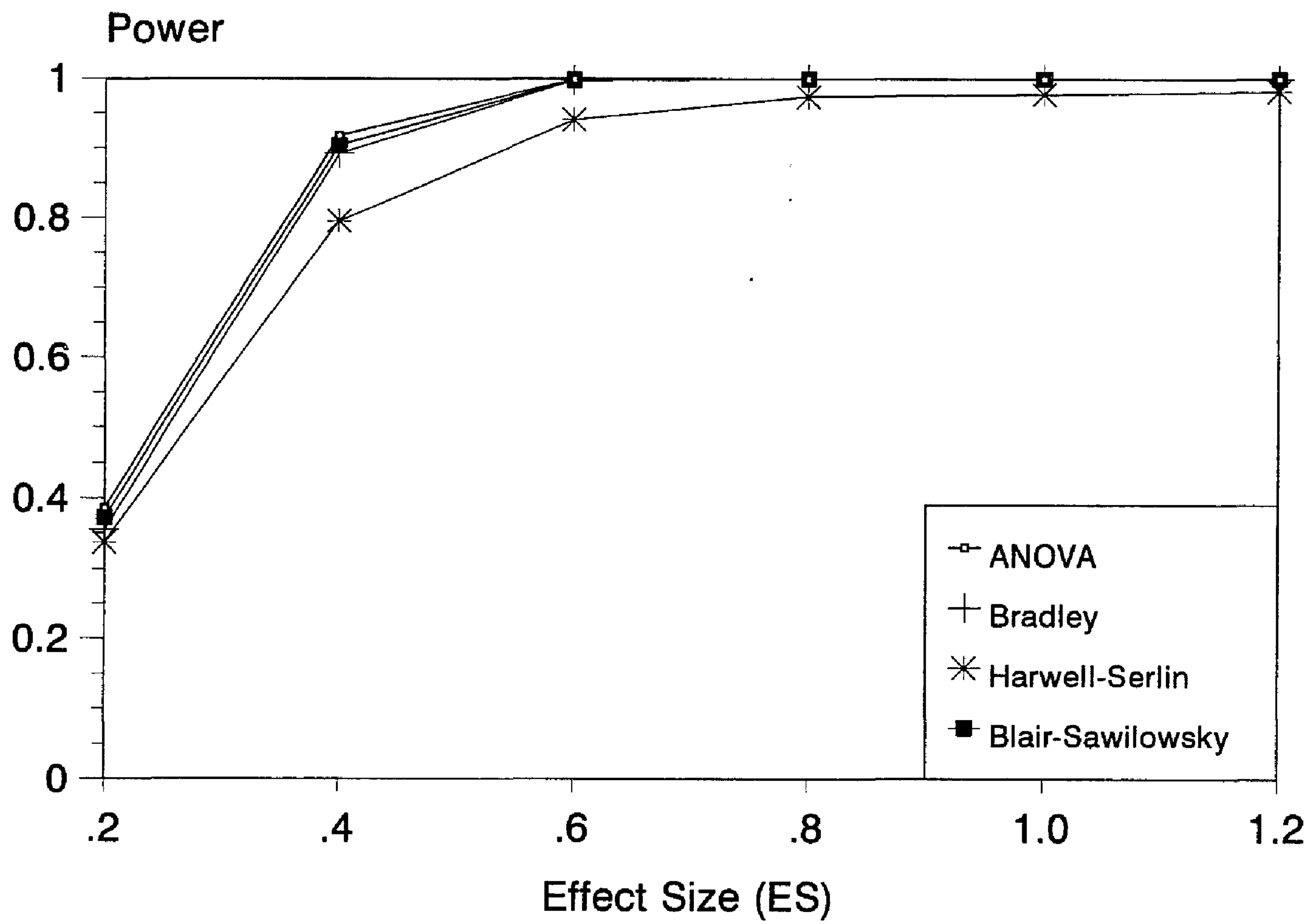

Figure 21. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the Gaussian distribution, alpha=.05 and n=35.

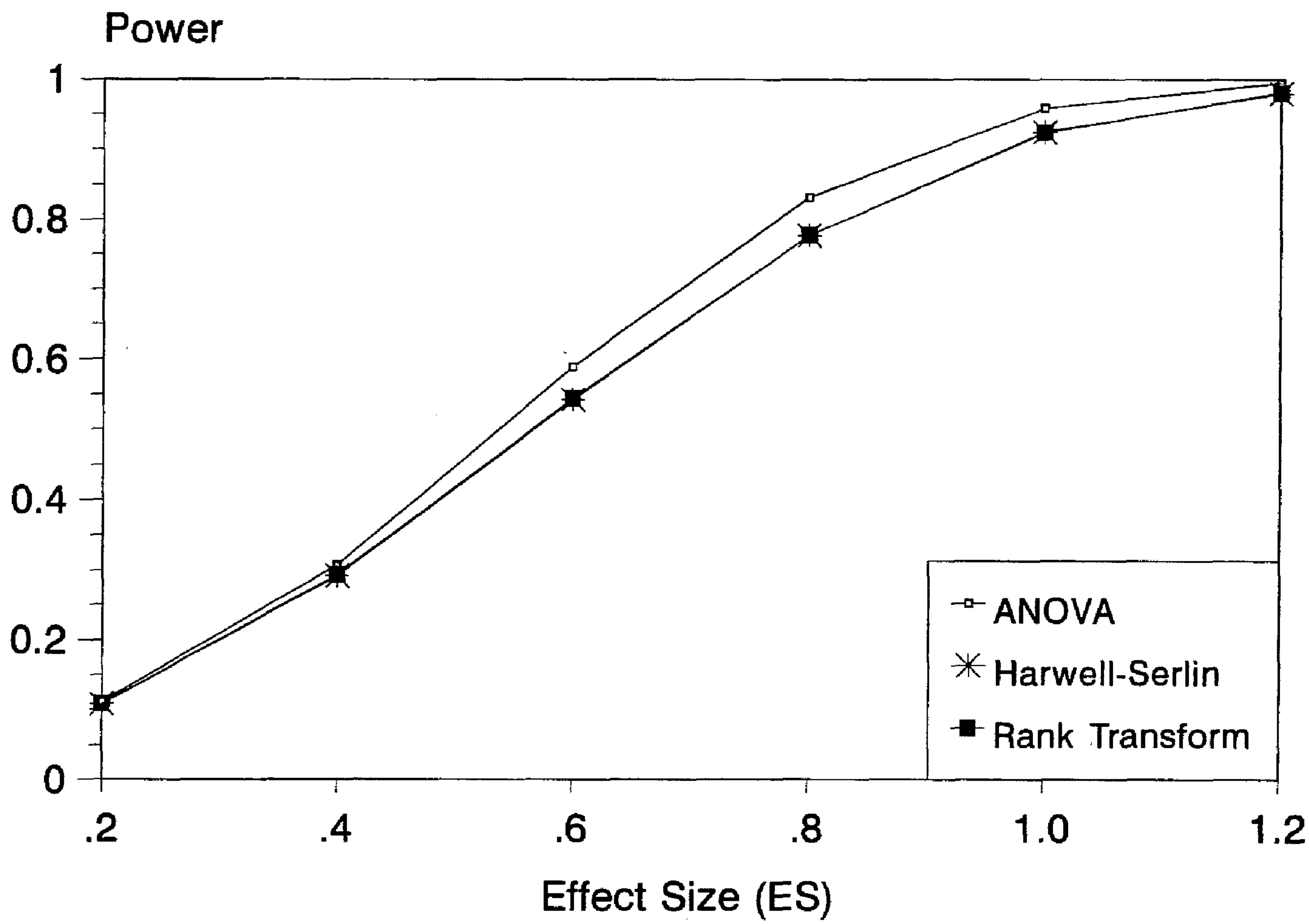

Figure 22. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the uniform distribution, alpha=.05 and n=7.

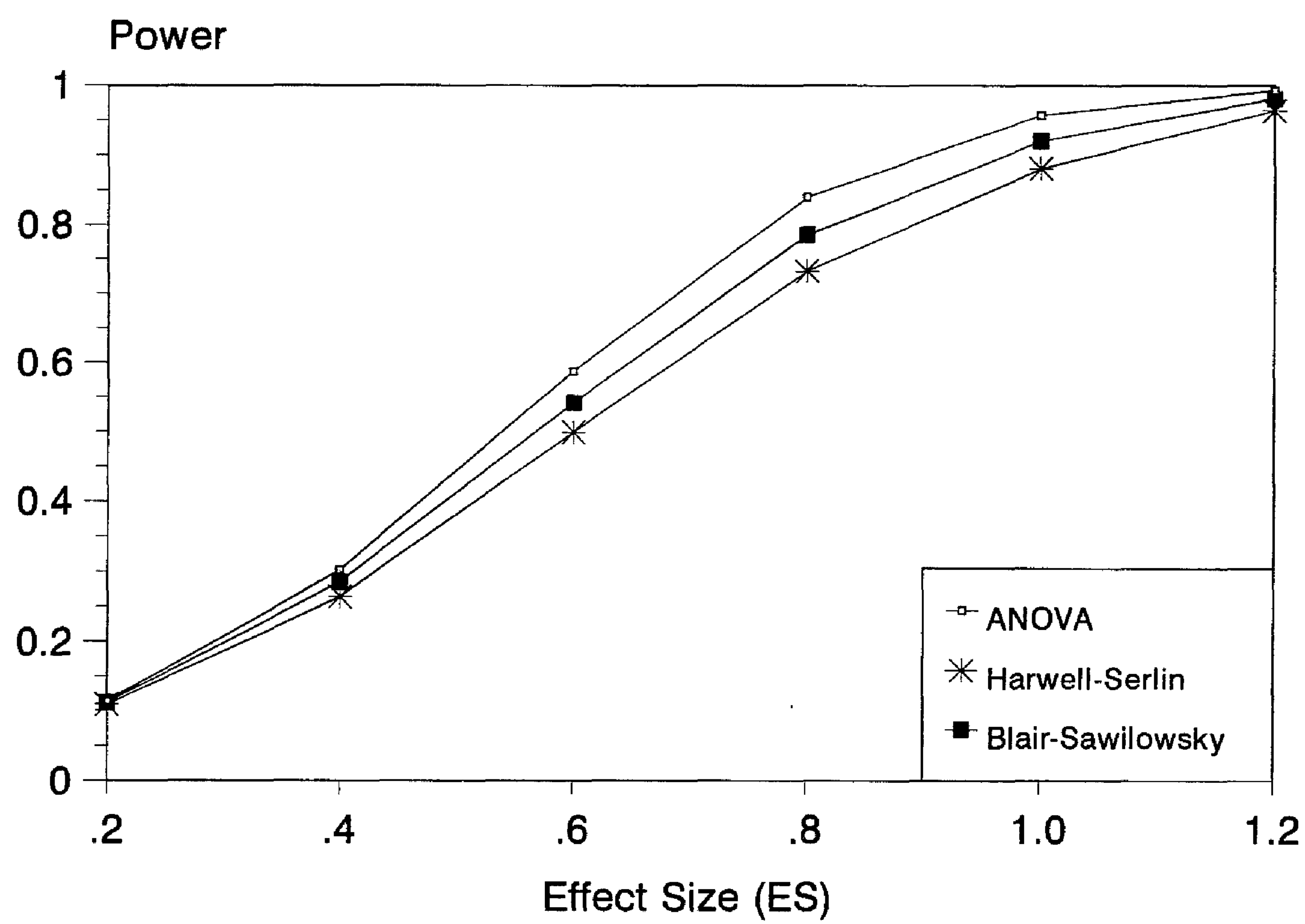

Figure 23. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=7.

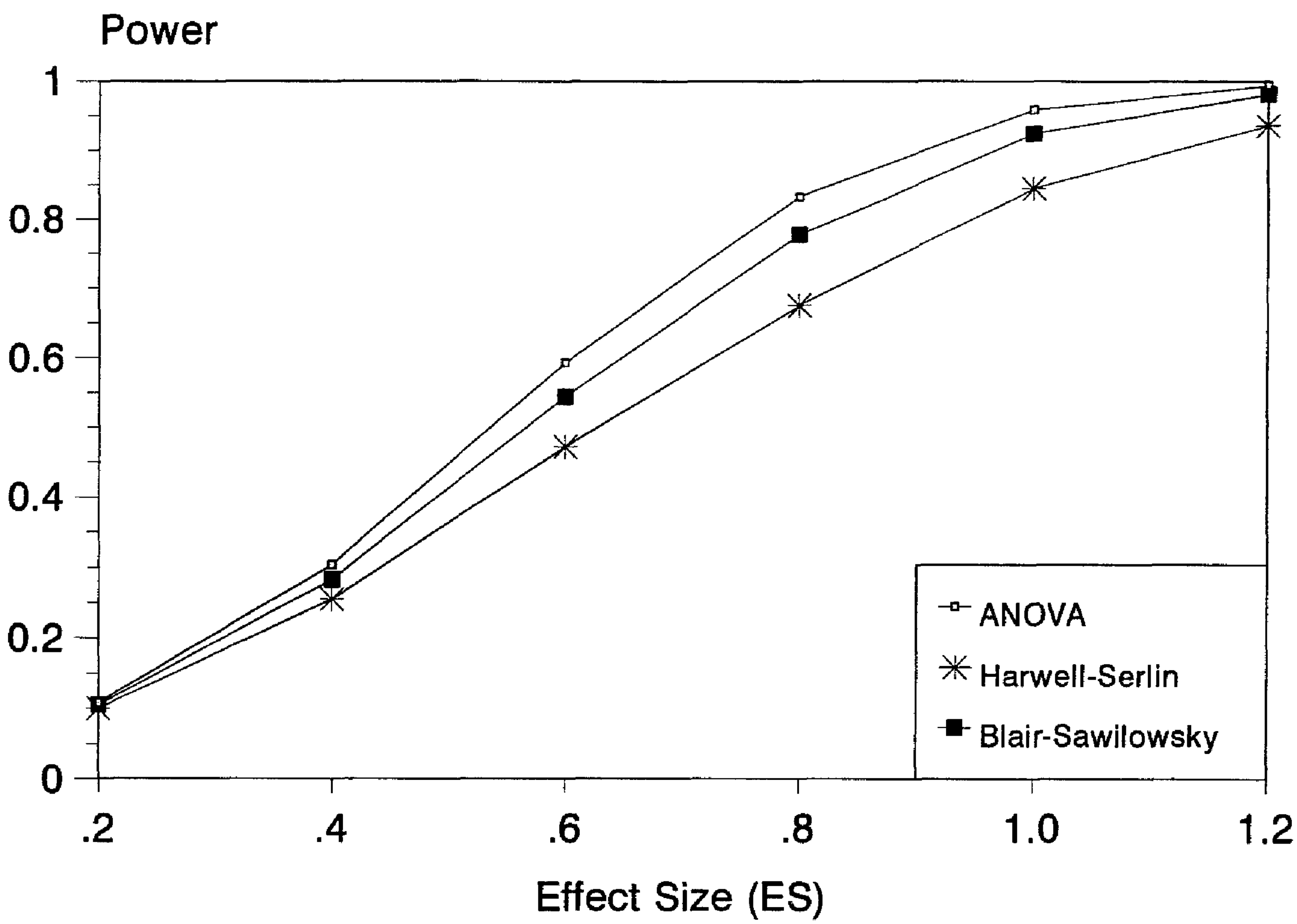

Figure 24. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=7.

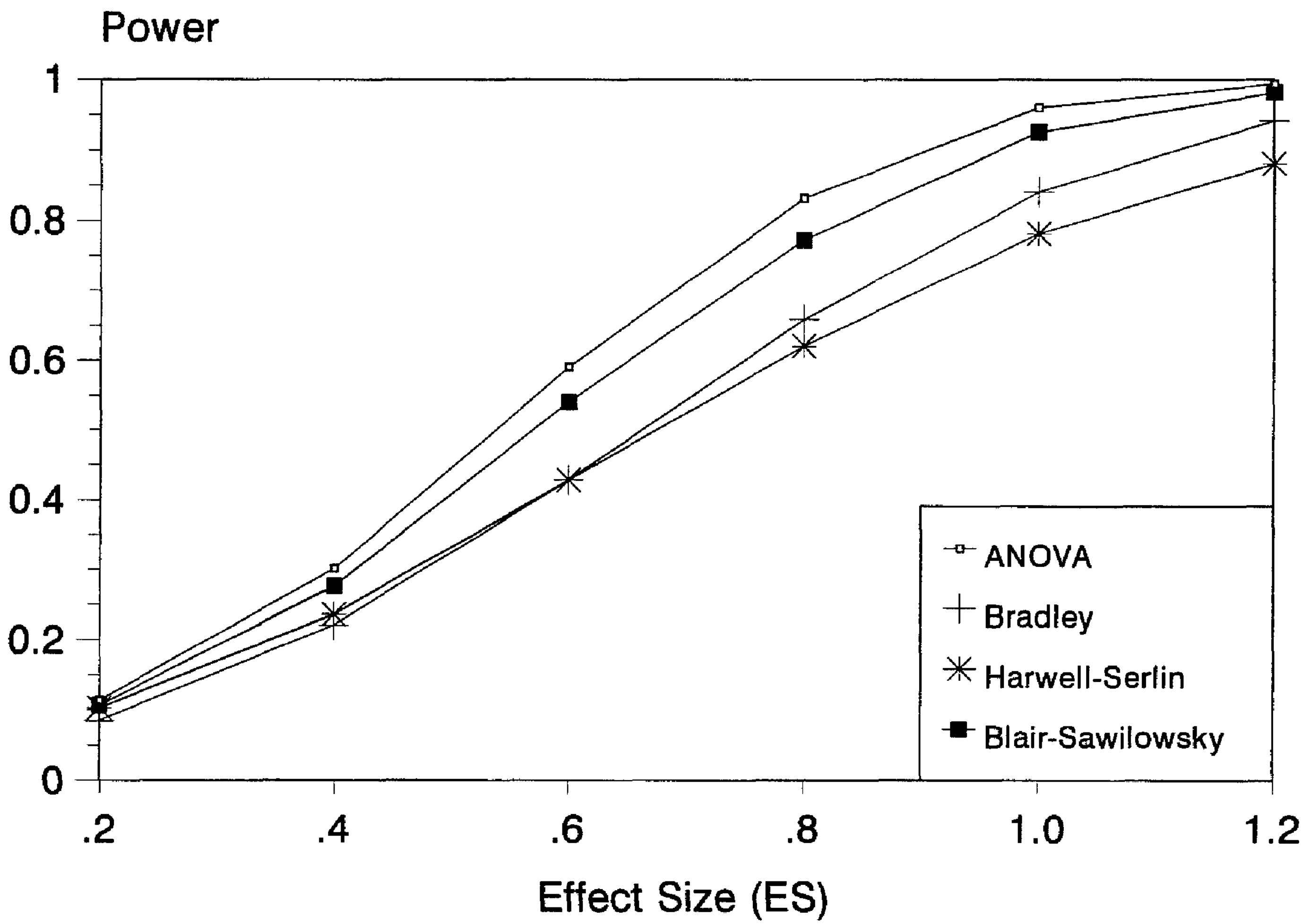

Figure 25. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=7.

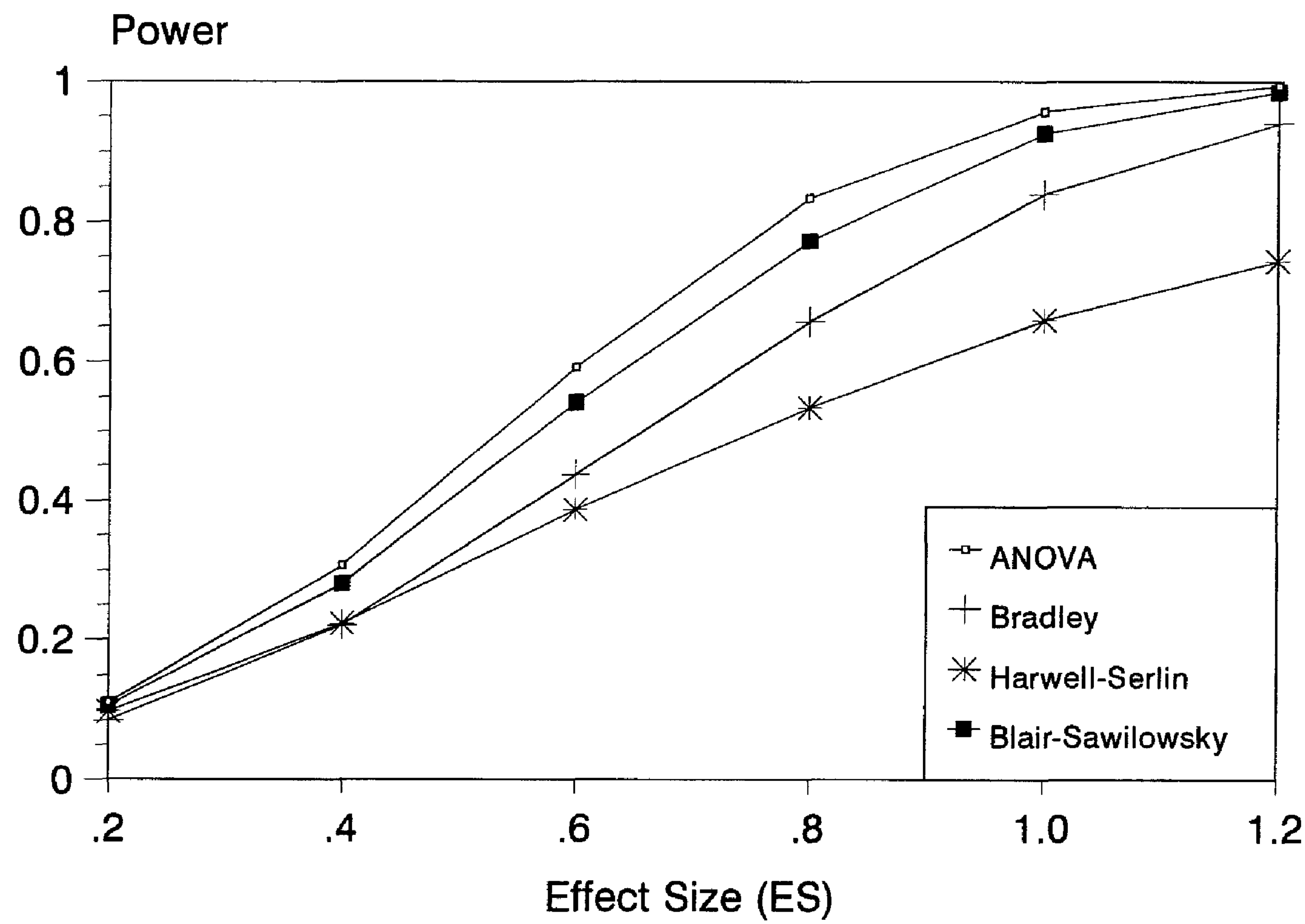

Figure 26. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=7.

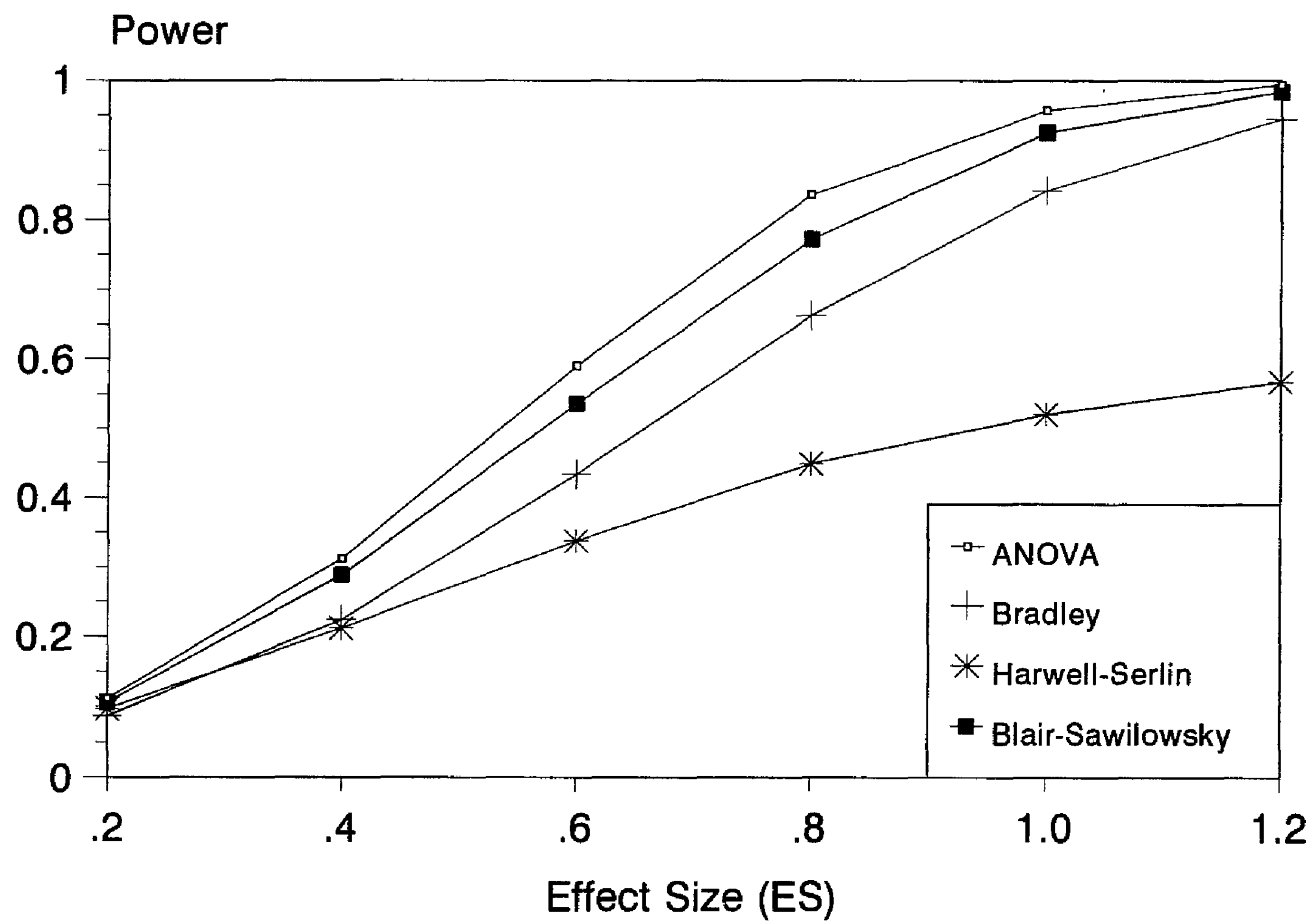

Figure 27. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=7.

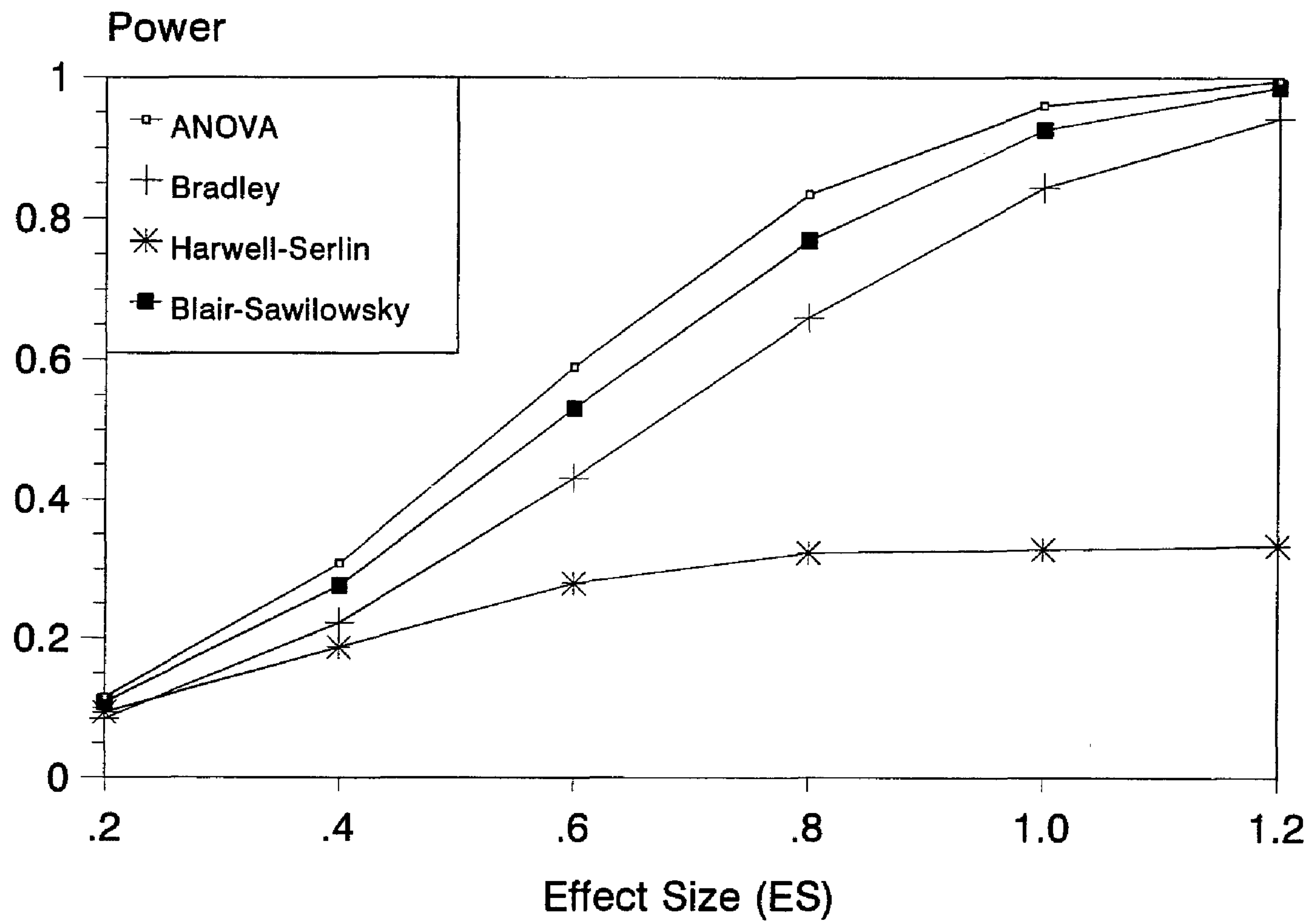

Figure 28. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=7.

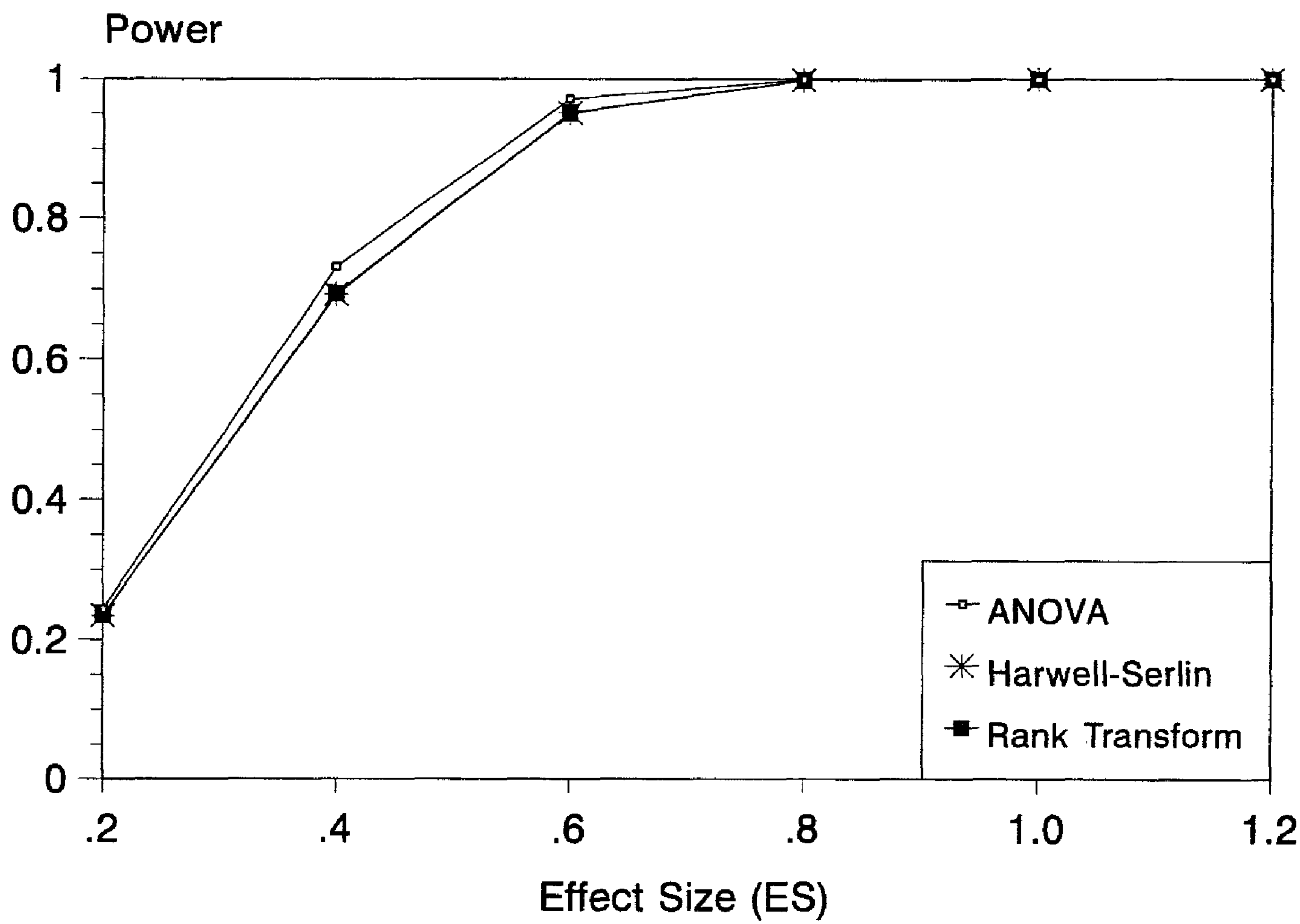

Figure 29. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the uniform distribution, alpha=.05, and n=21.

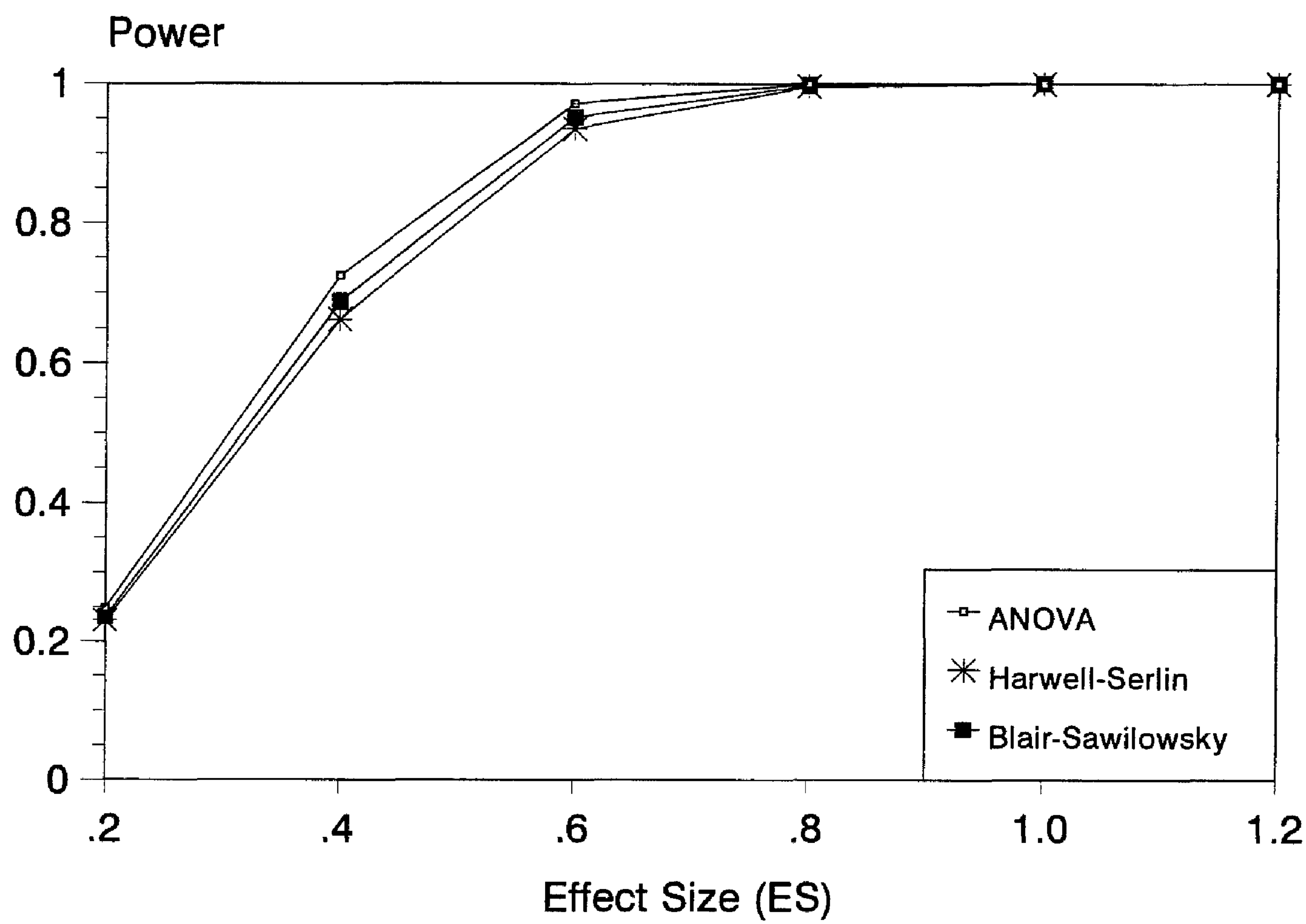

Figure 30. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=21.

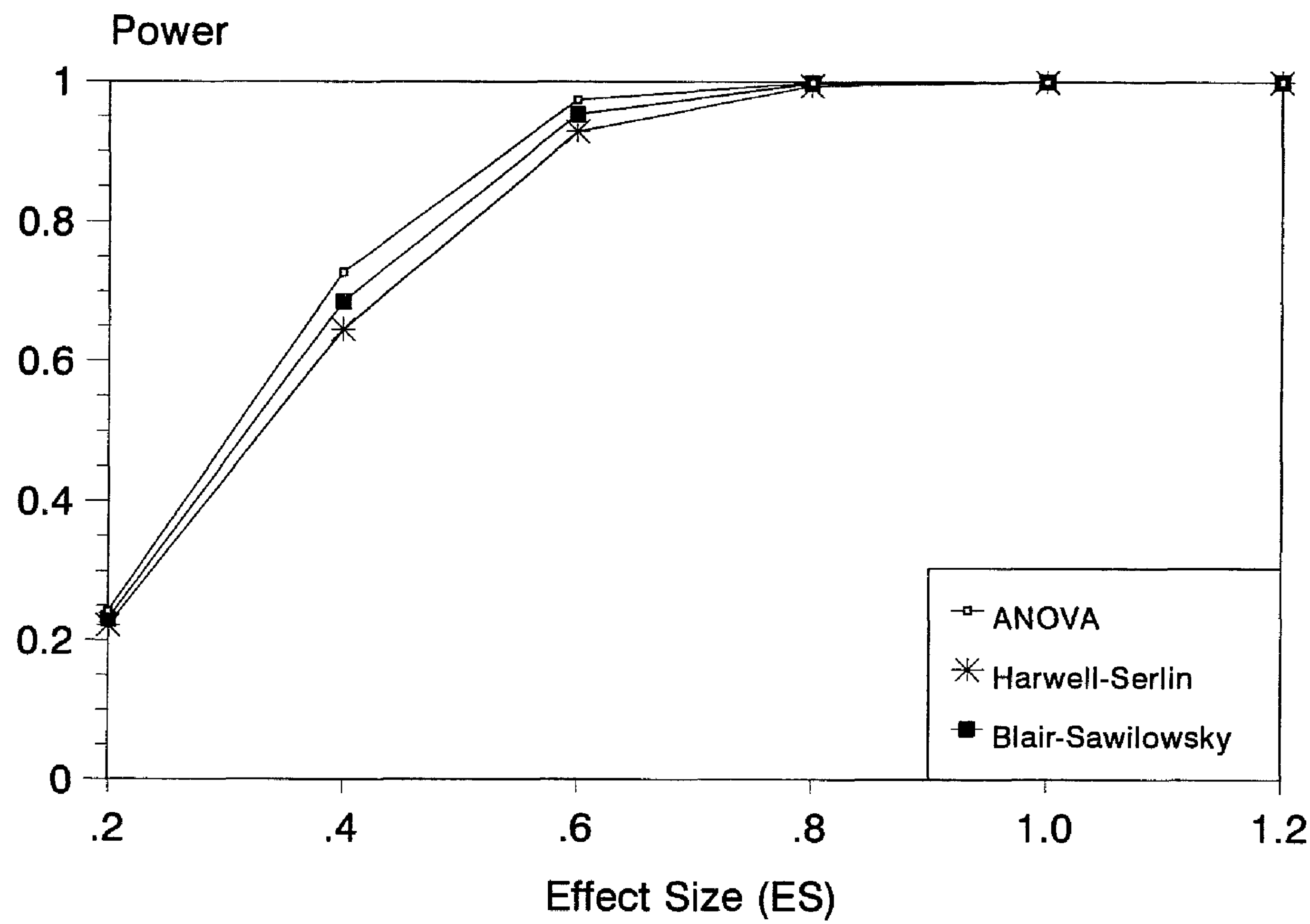

Figure 31. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=21.

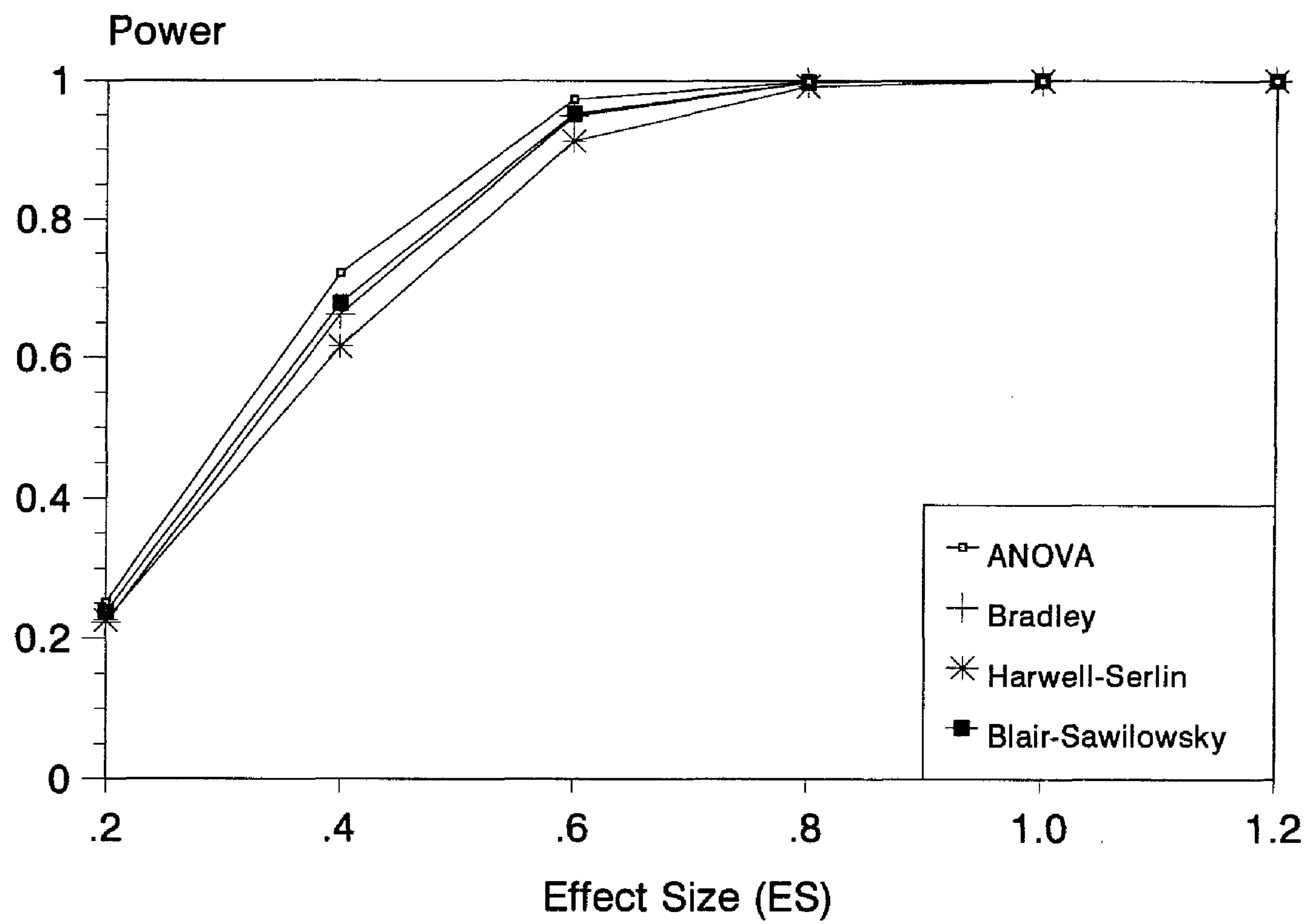

Figure 32. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=21.

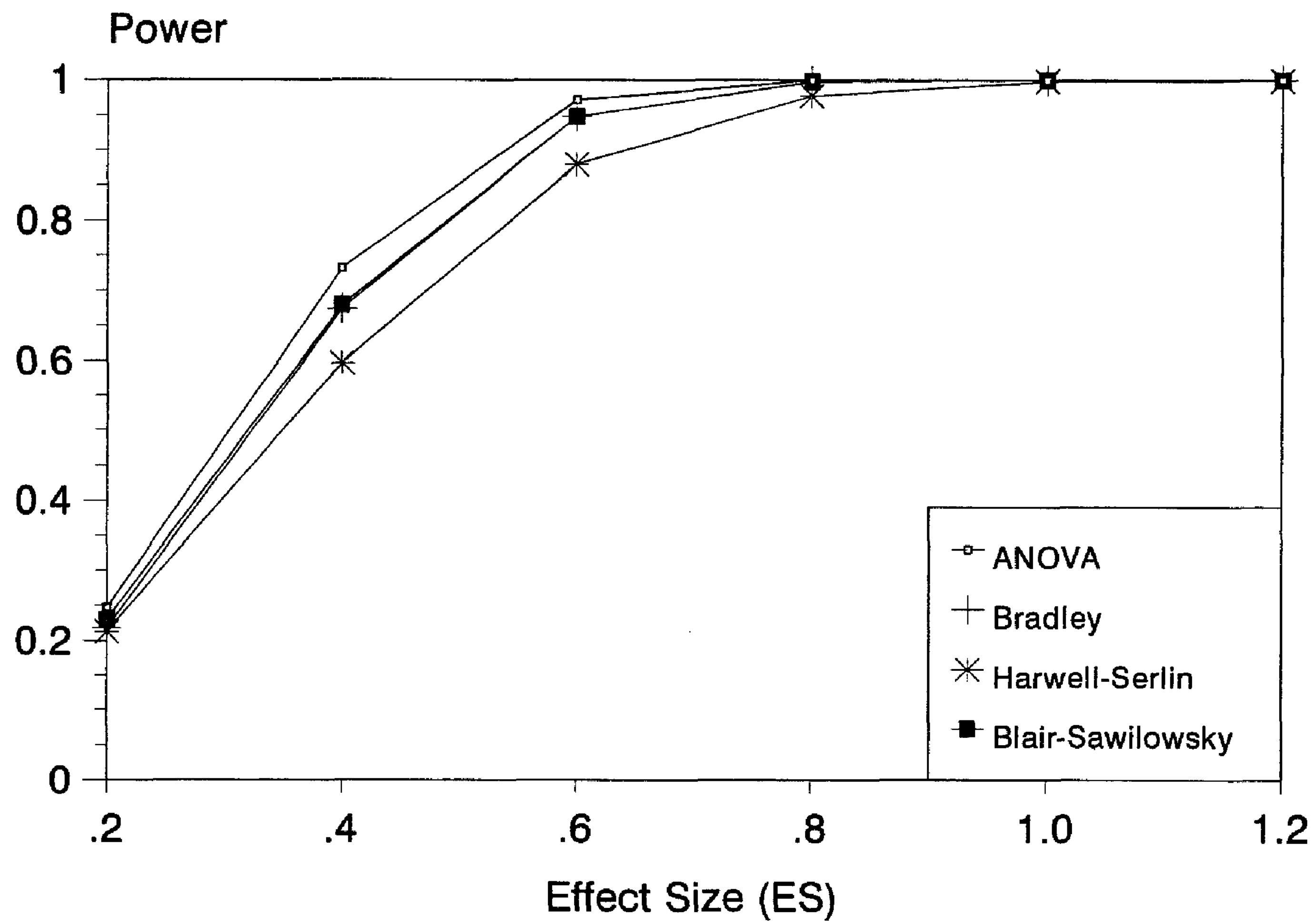

Figure 33. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=21.

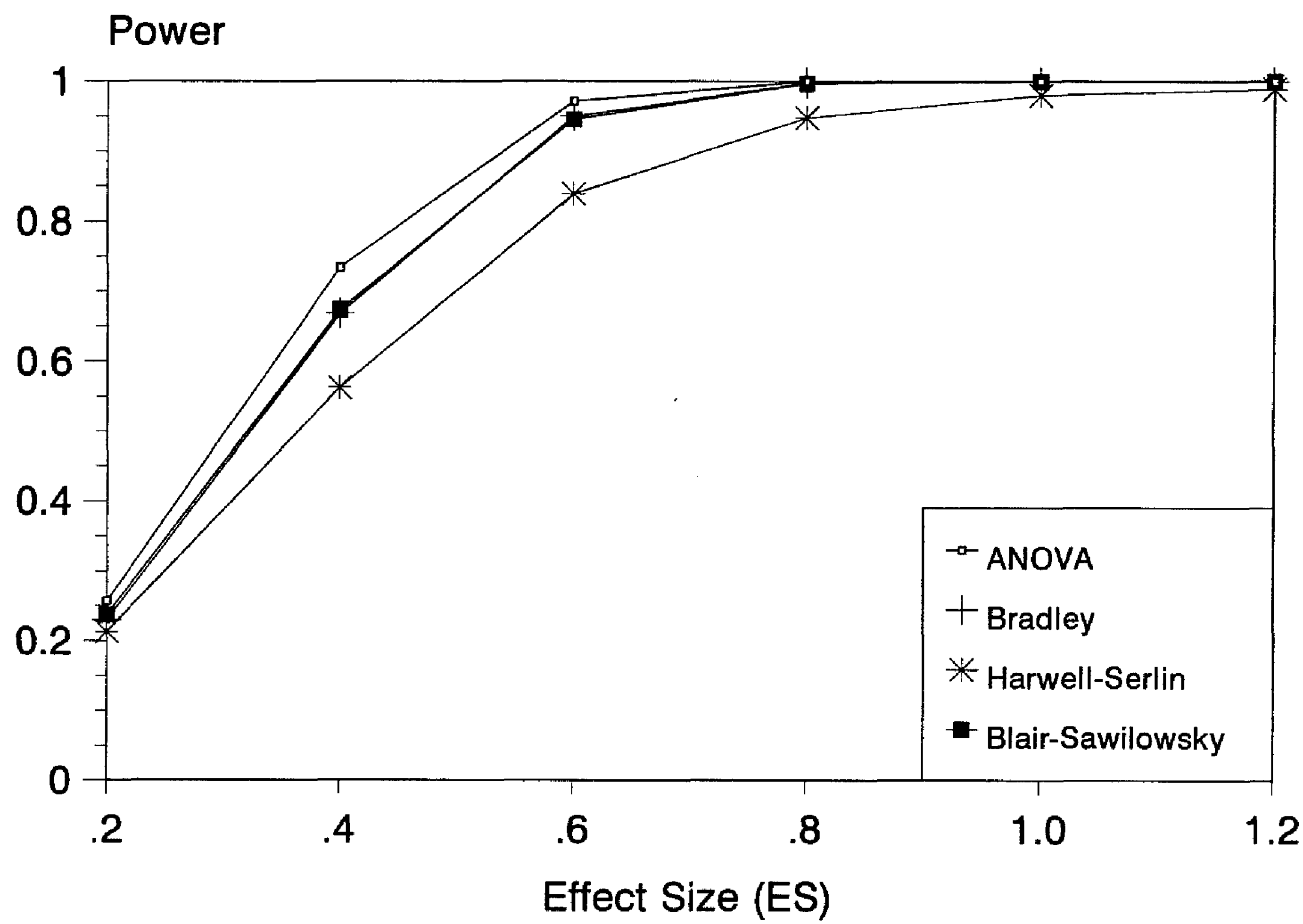

Figure 34. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=21.

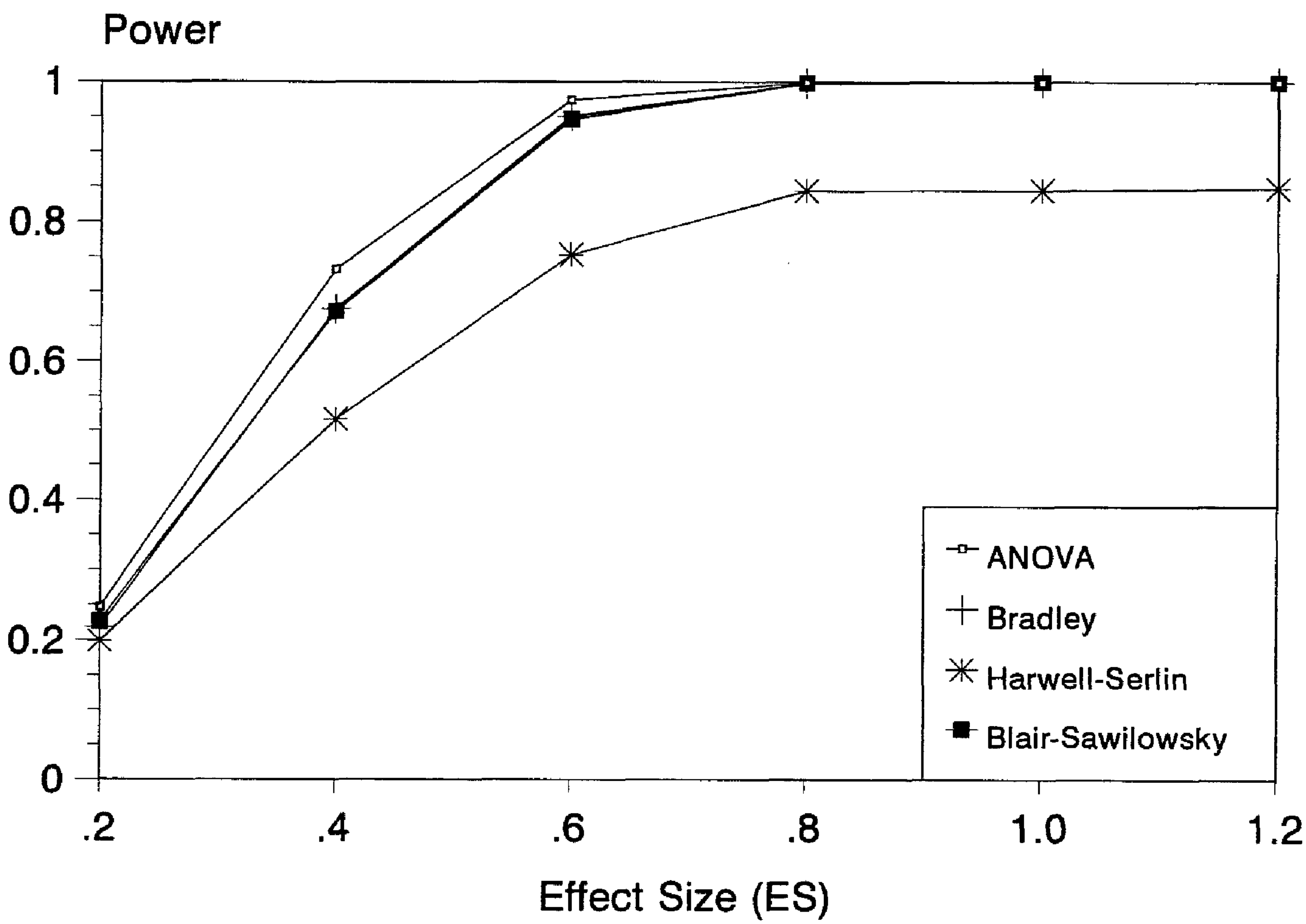

Figure 35. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=21.

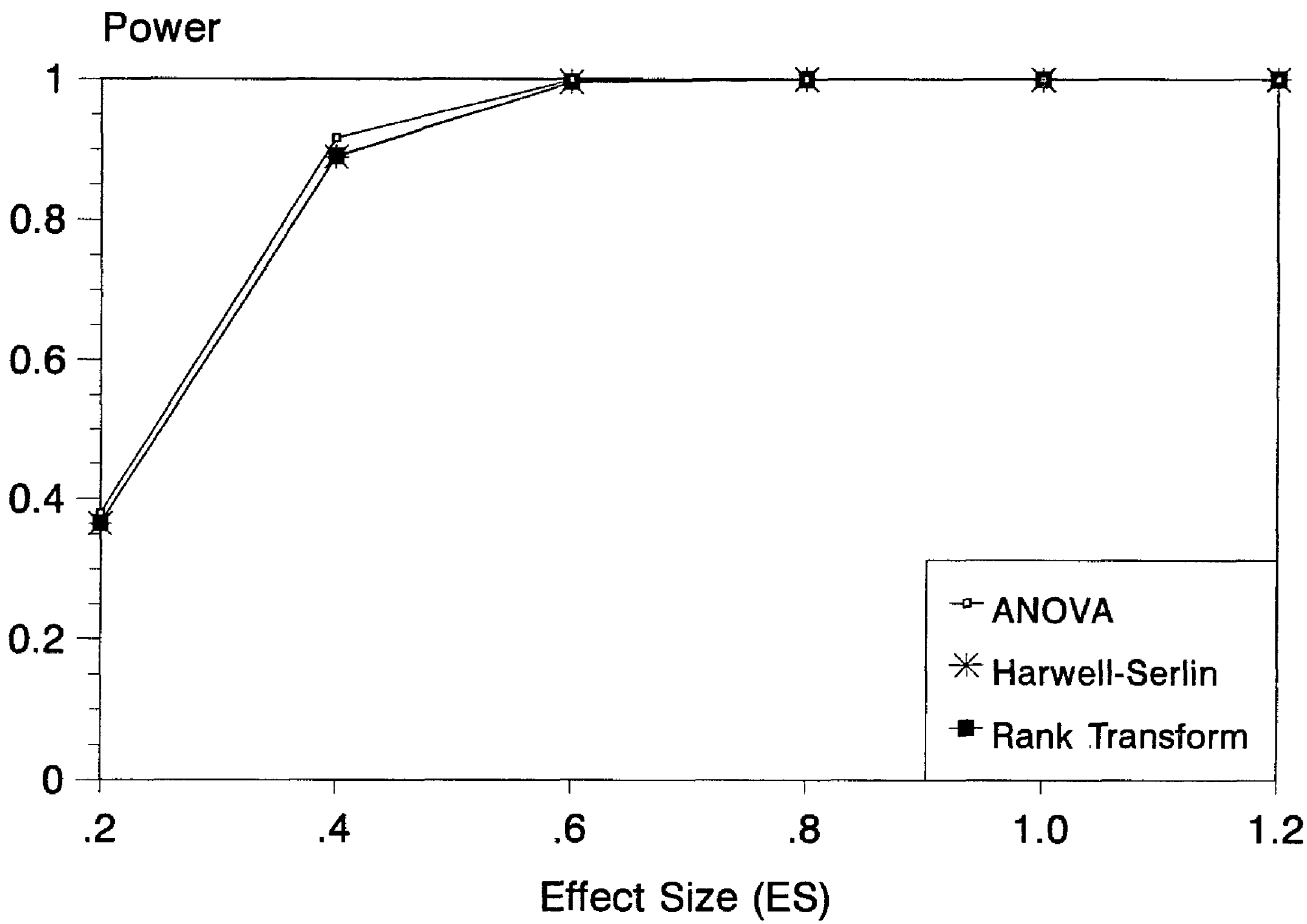

Figure 36. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the uniform distribution, alpha=.05 and n=35.

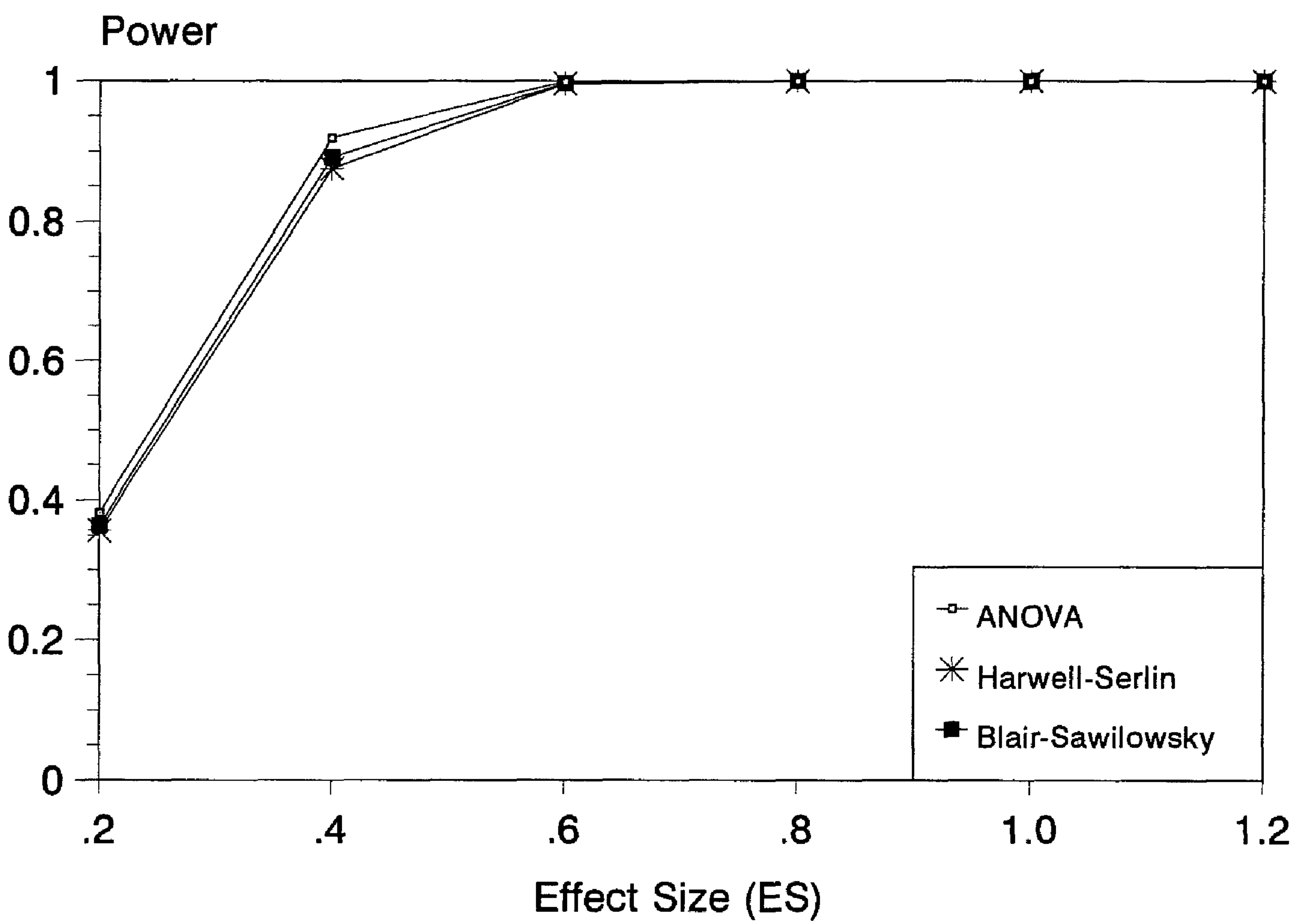

Figure 37. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=35.

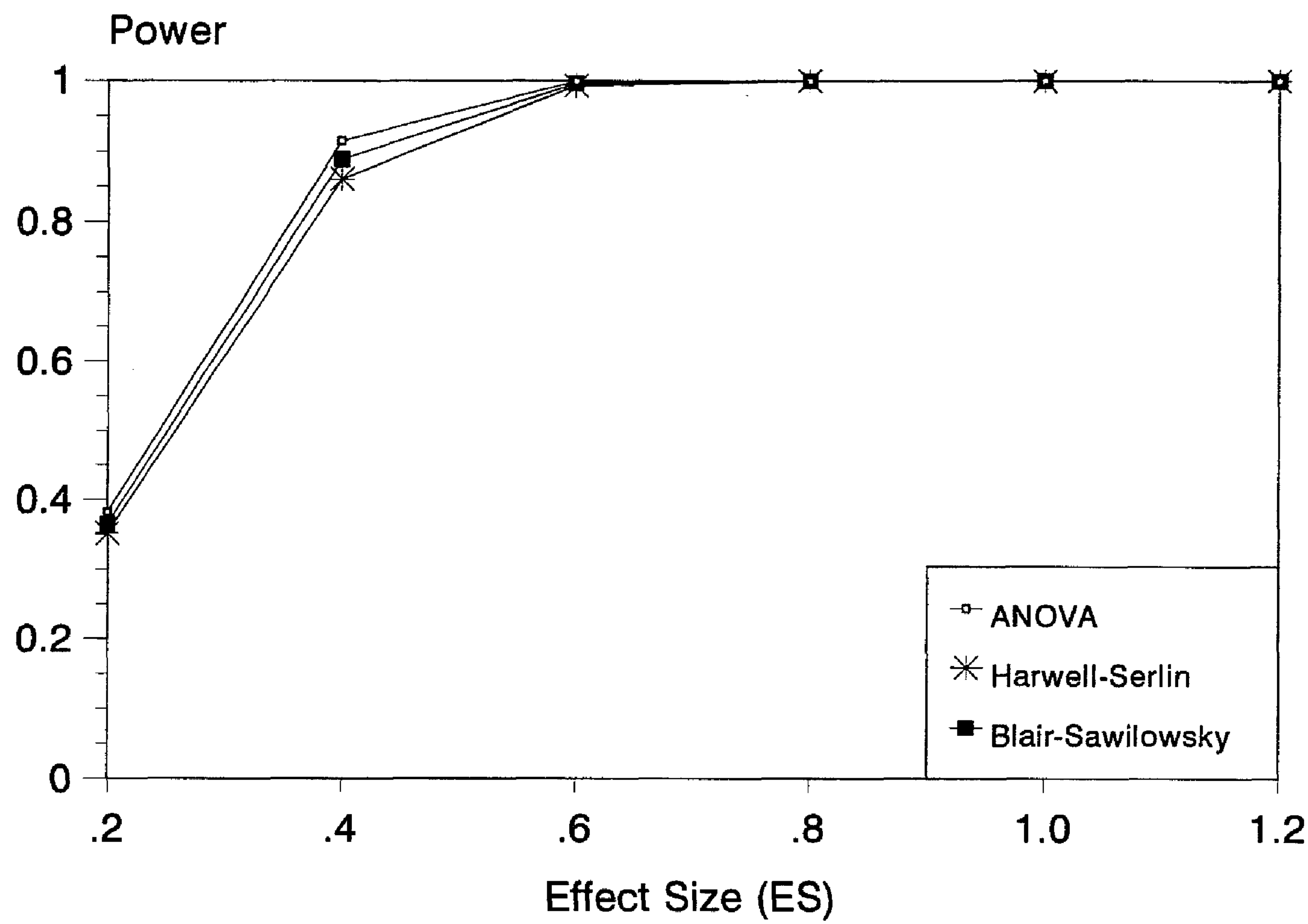

Figure 38. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=35.

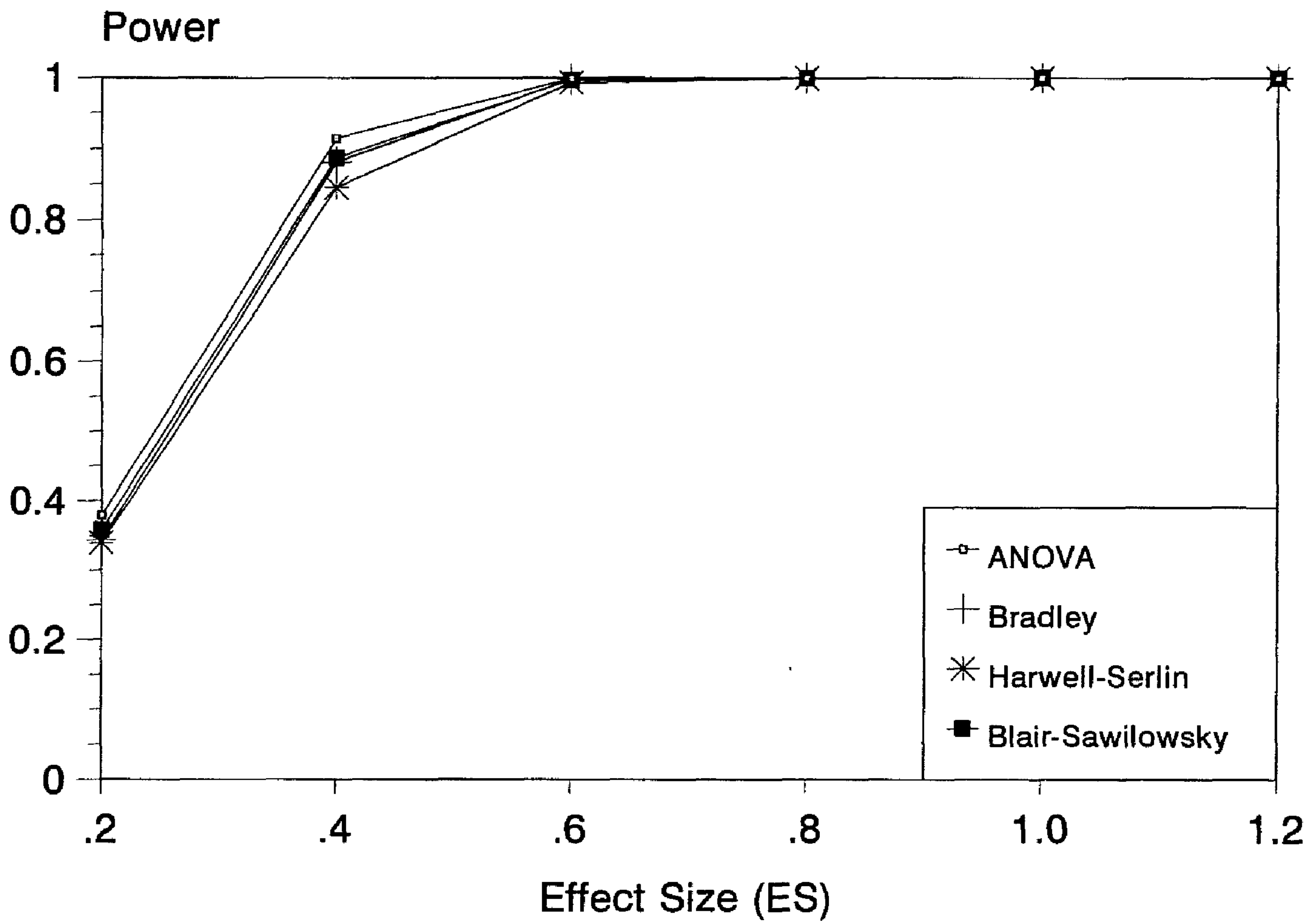

Figure 39. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=35.

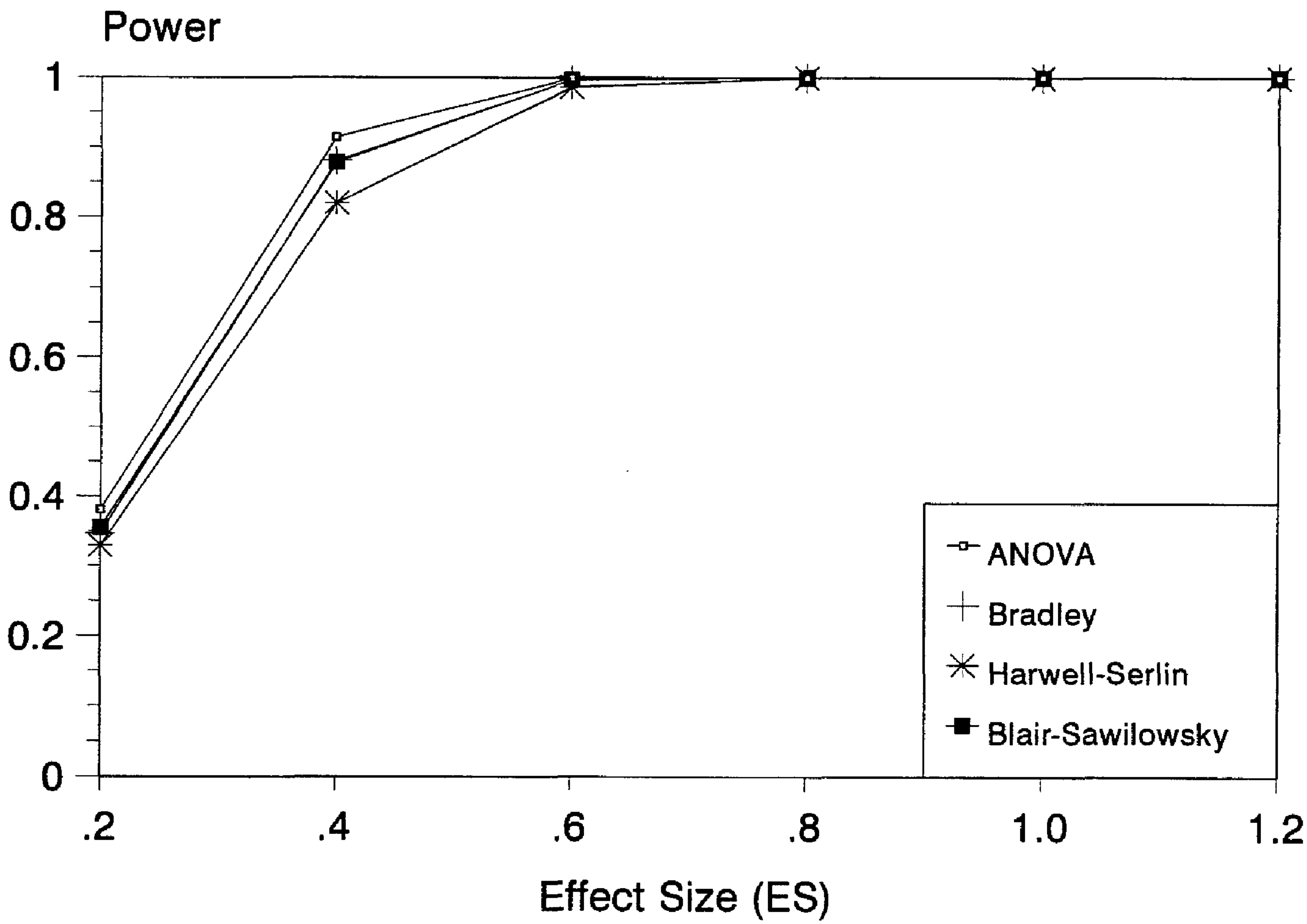

Figure 40. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=35.

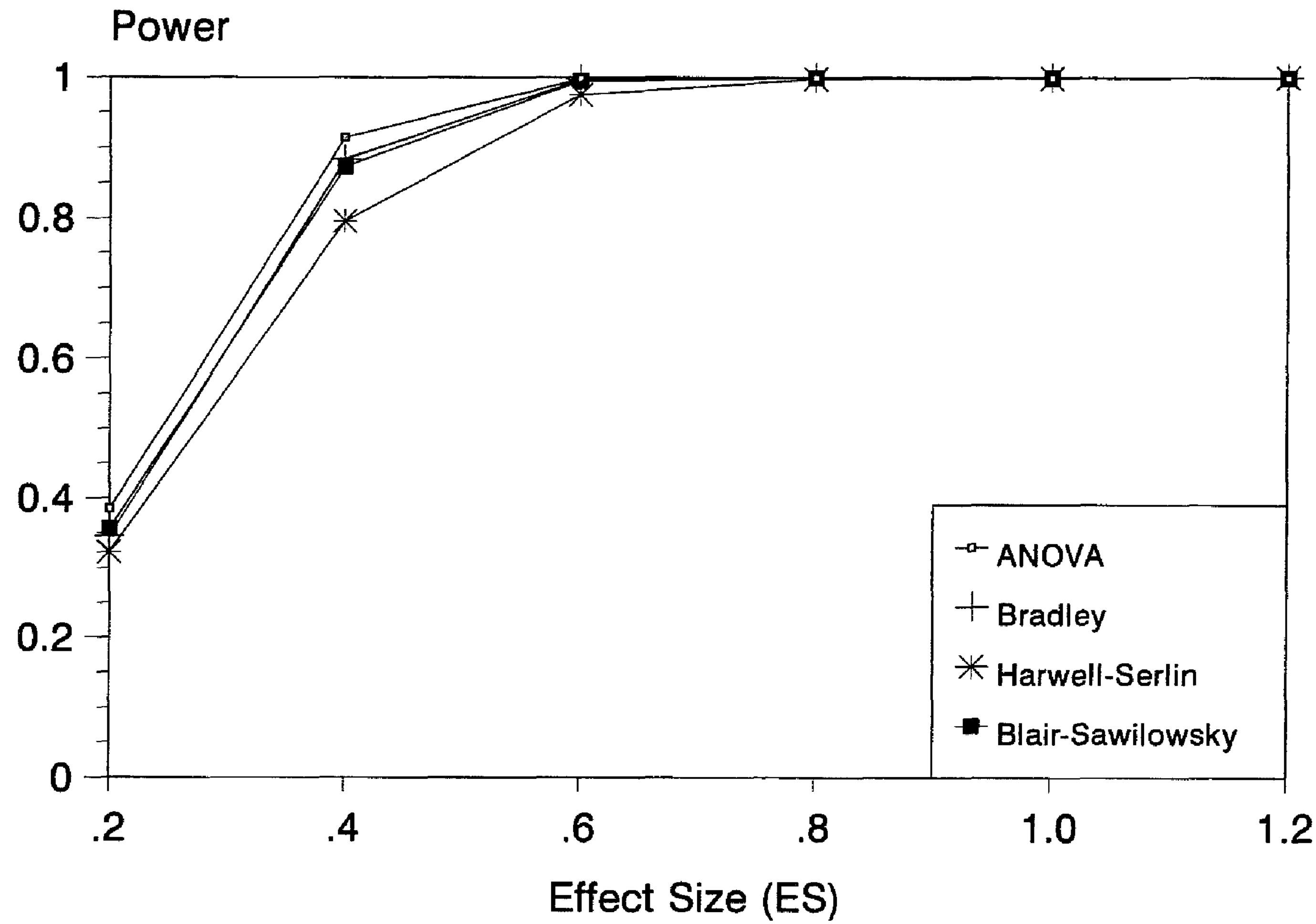

Figure 41. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05, and n=35.

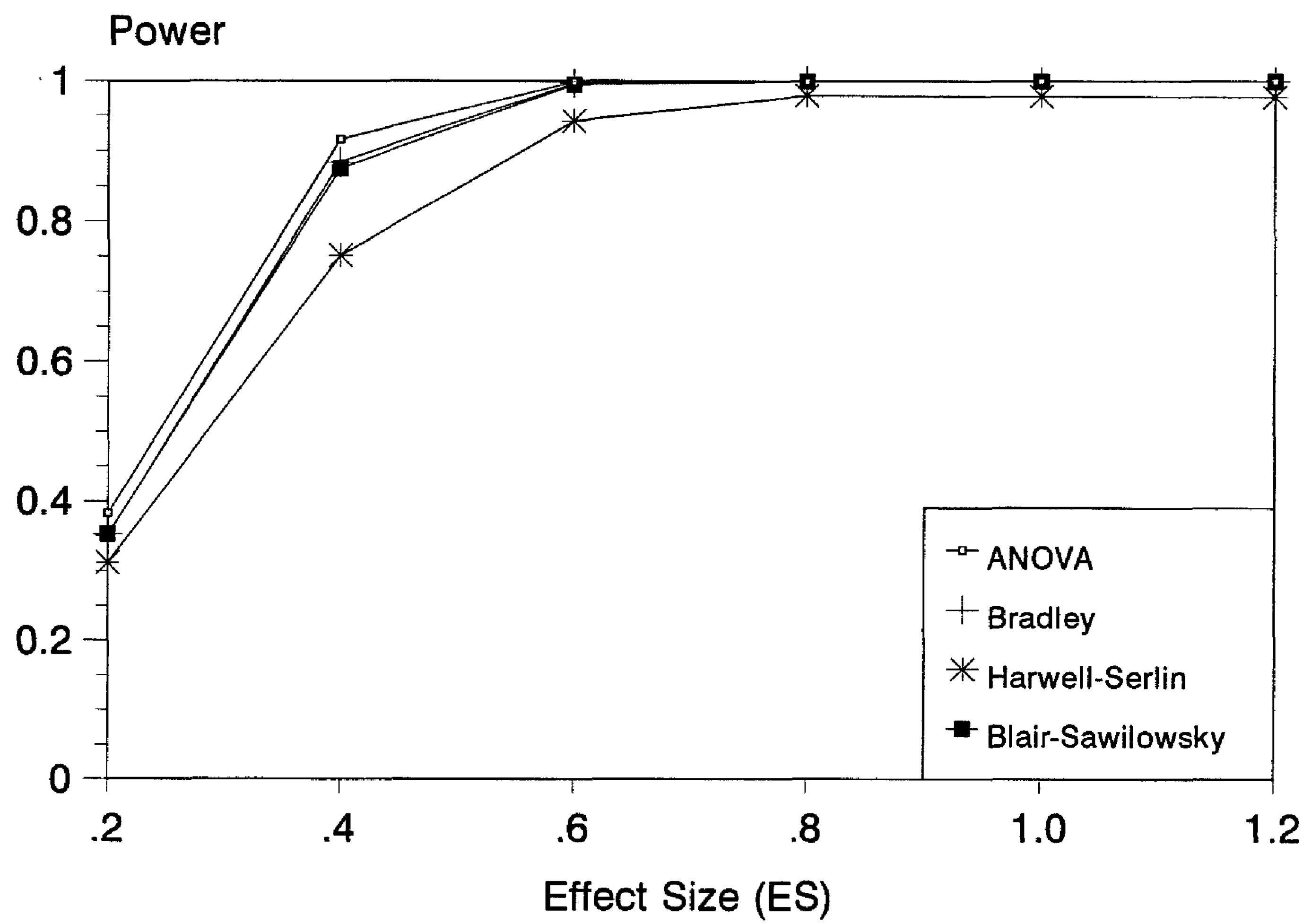

Figure 42. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the uniform distribution, alpha=.05 and n=35.

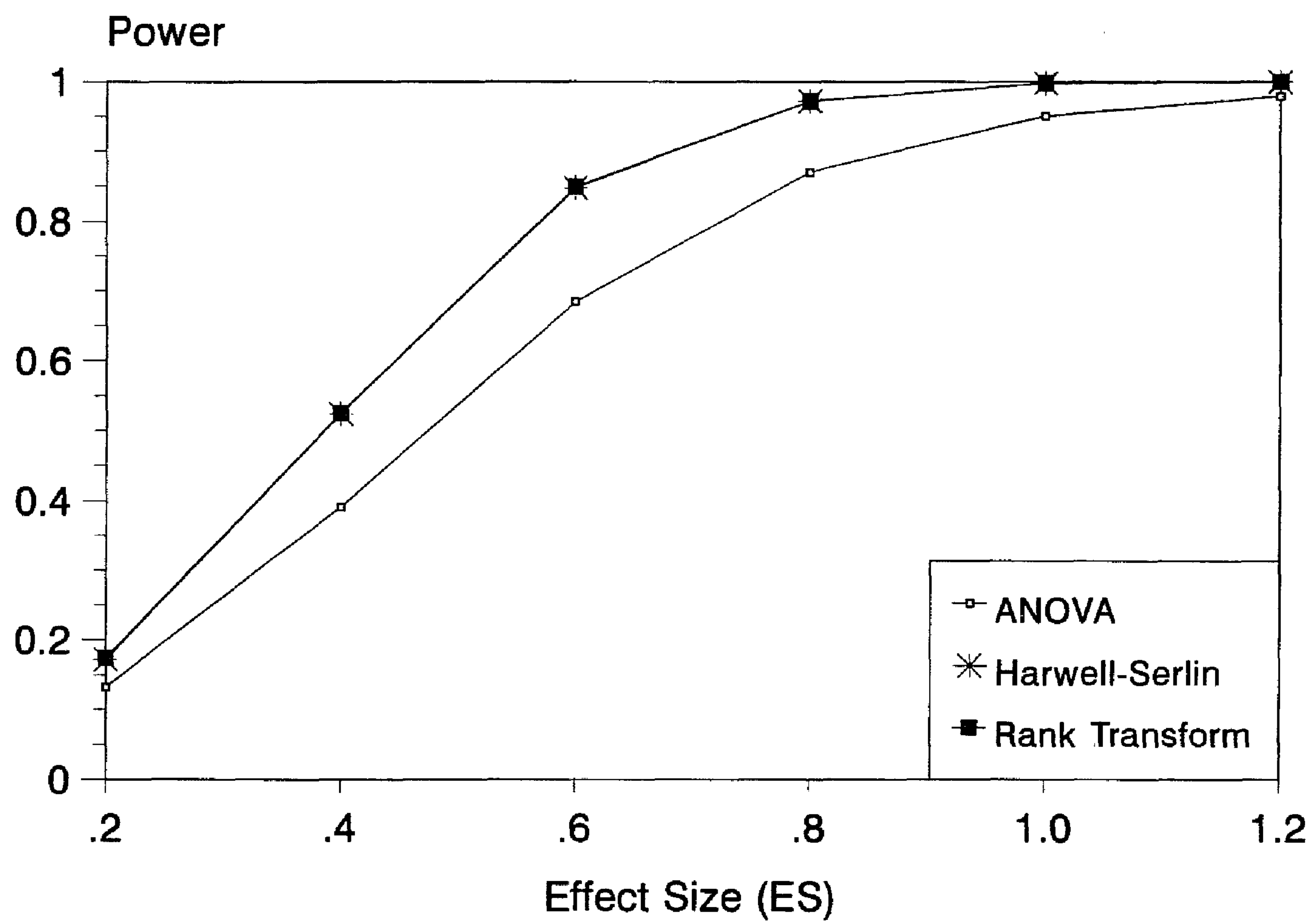

Figure 43. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=7.

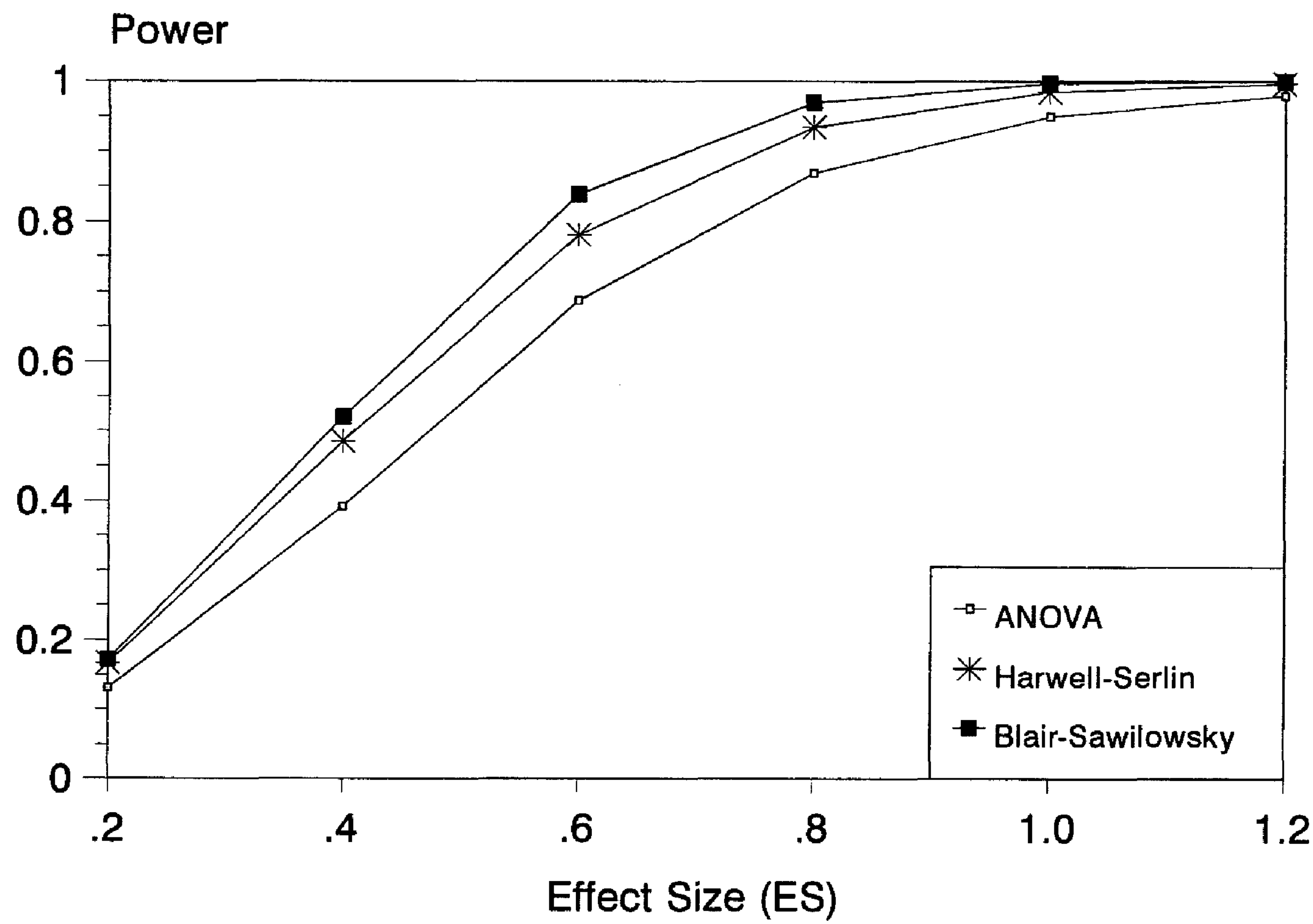

Figure 44. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=7.

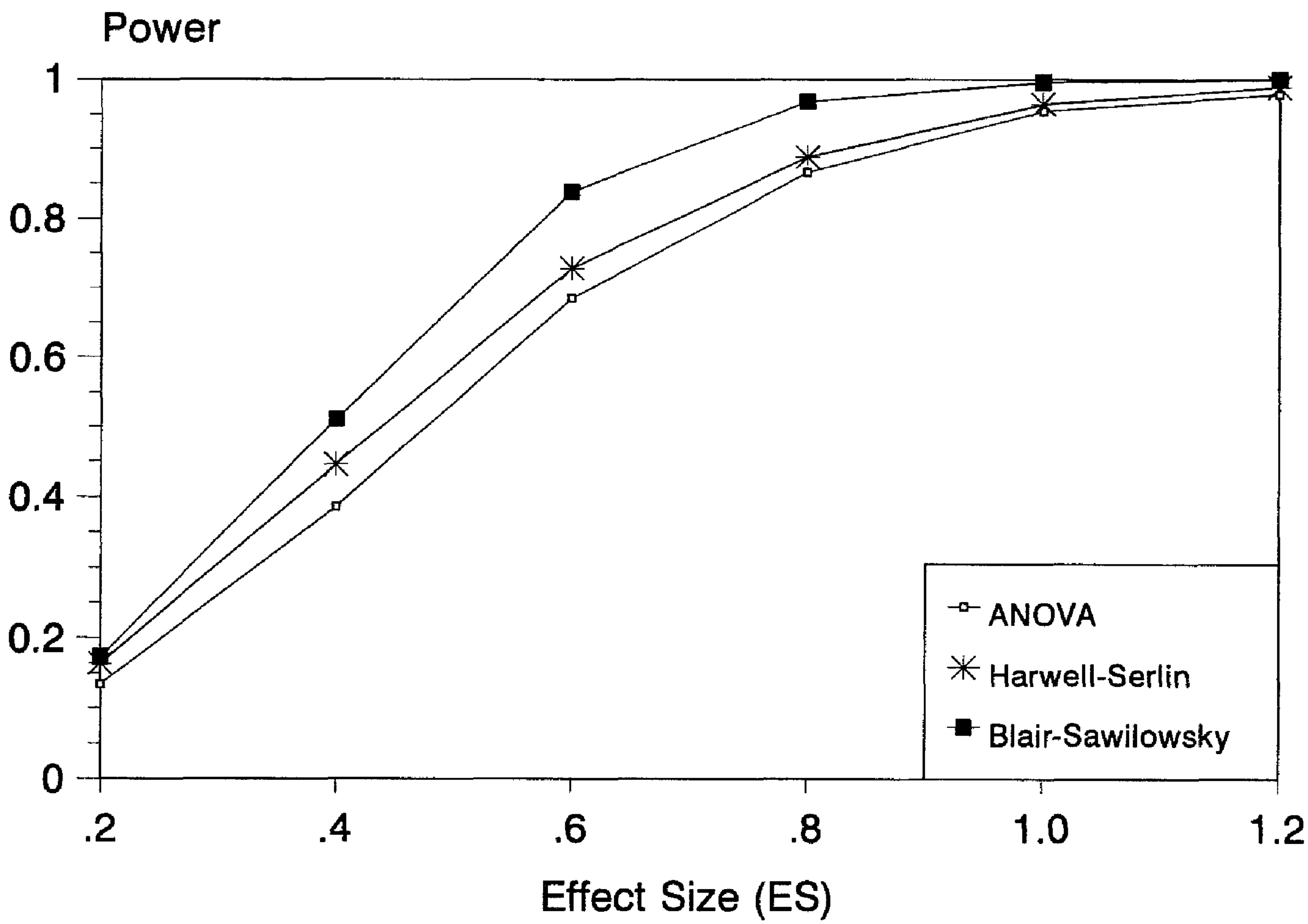

Figure 45. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=7.

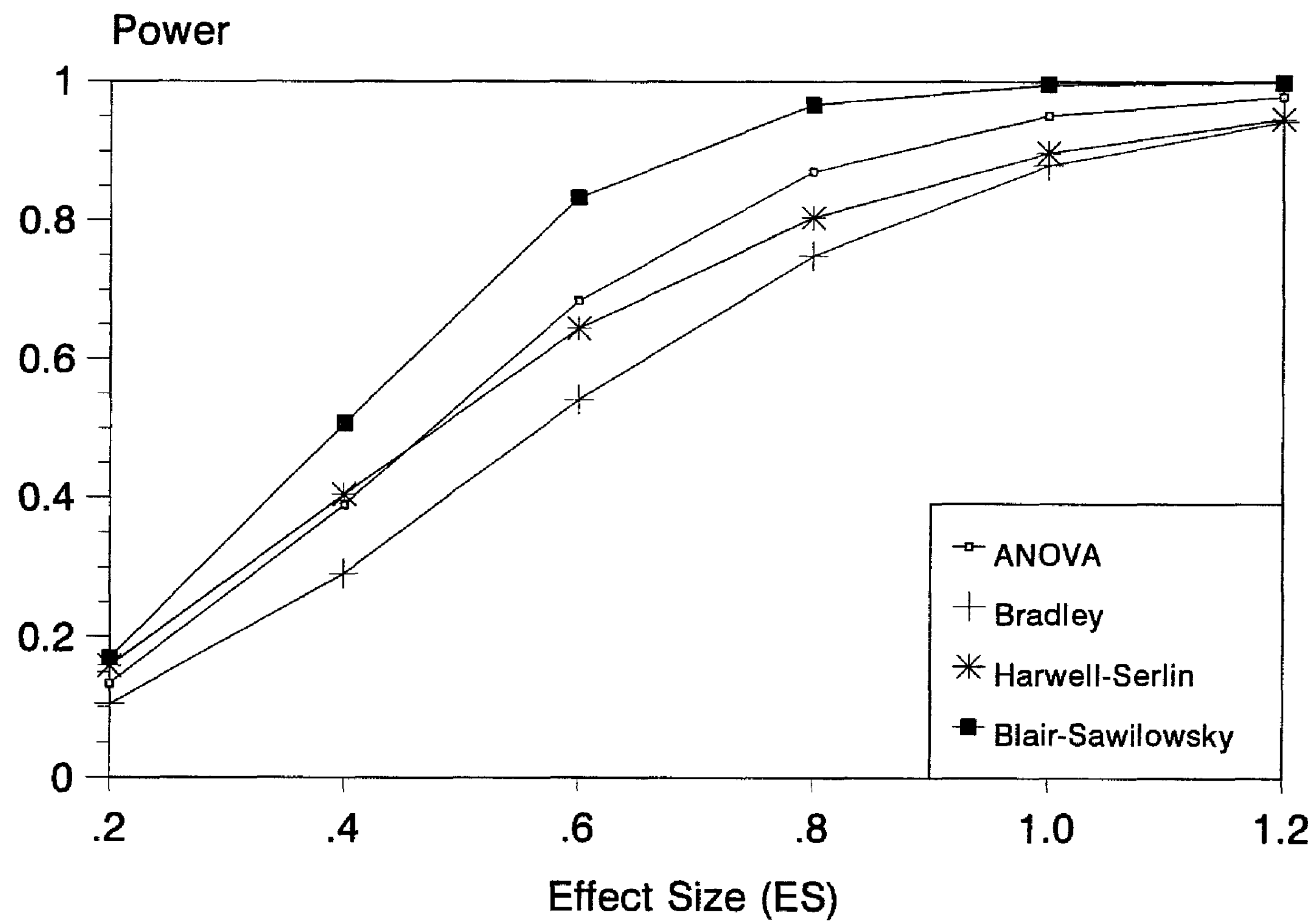

Figure 46. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=7.

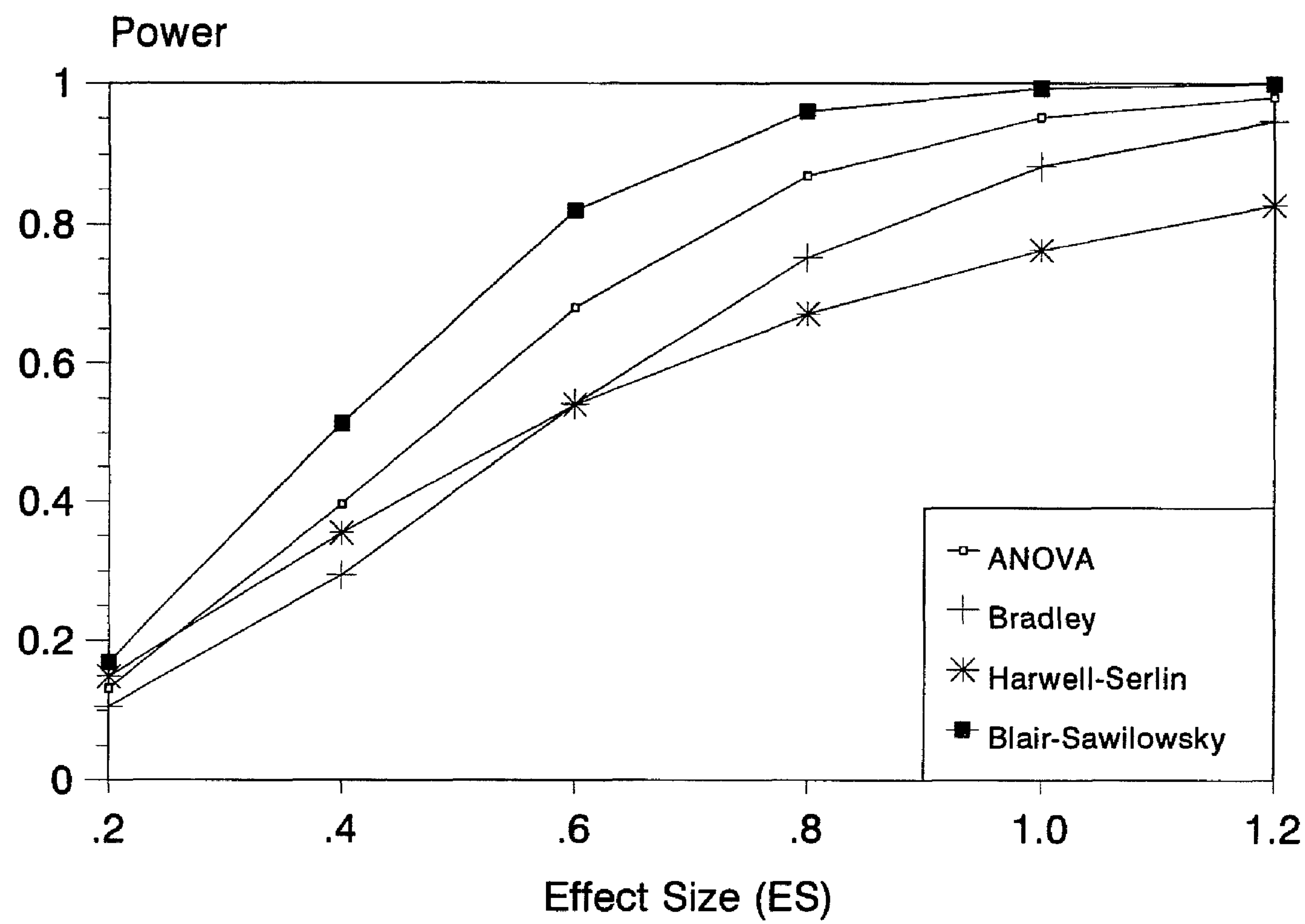

Figure 47. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=7.

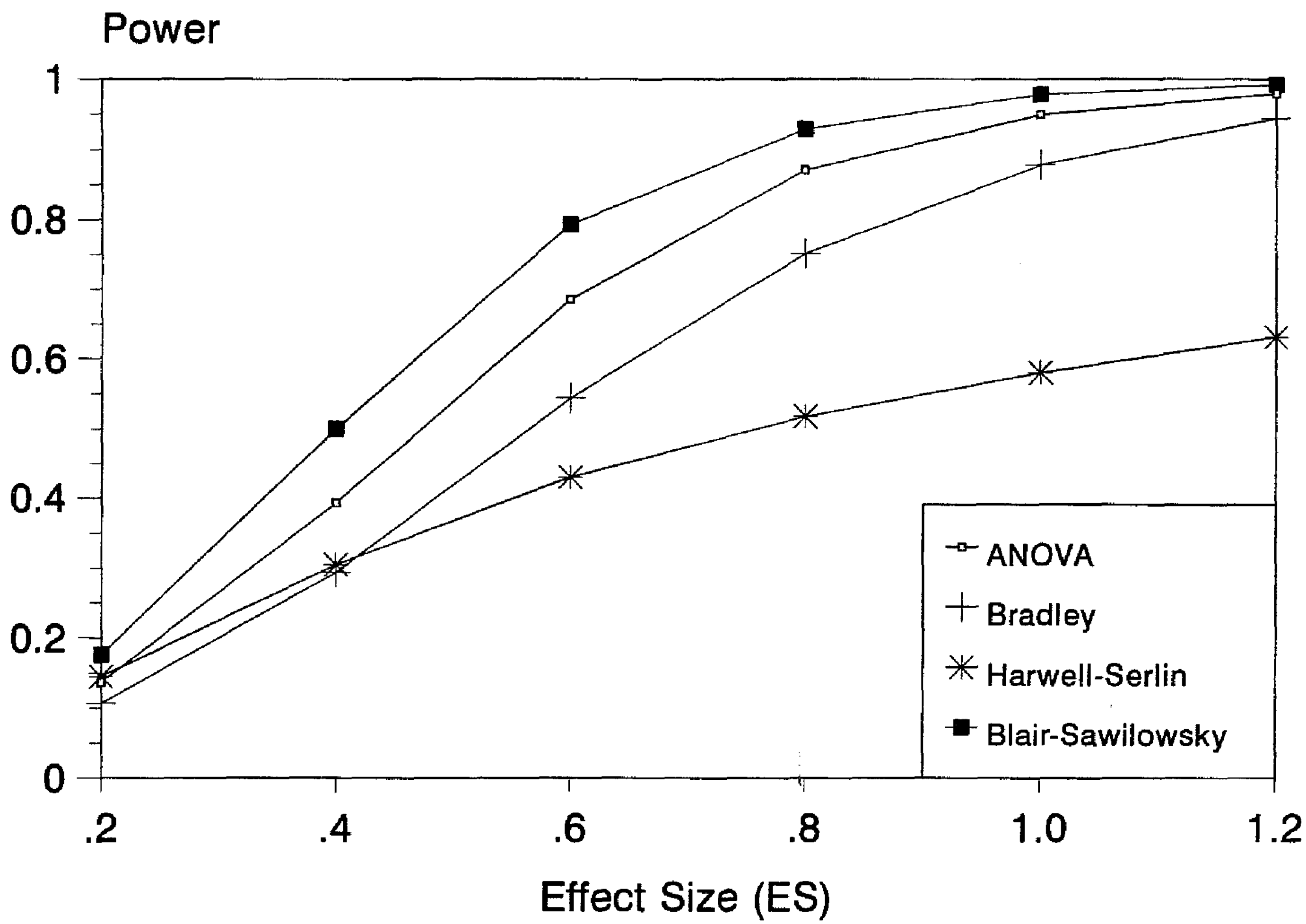

Figure 48. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=7.

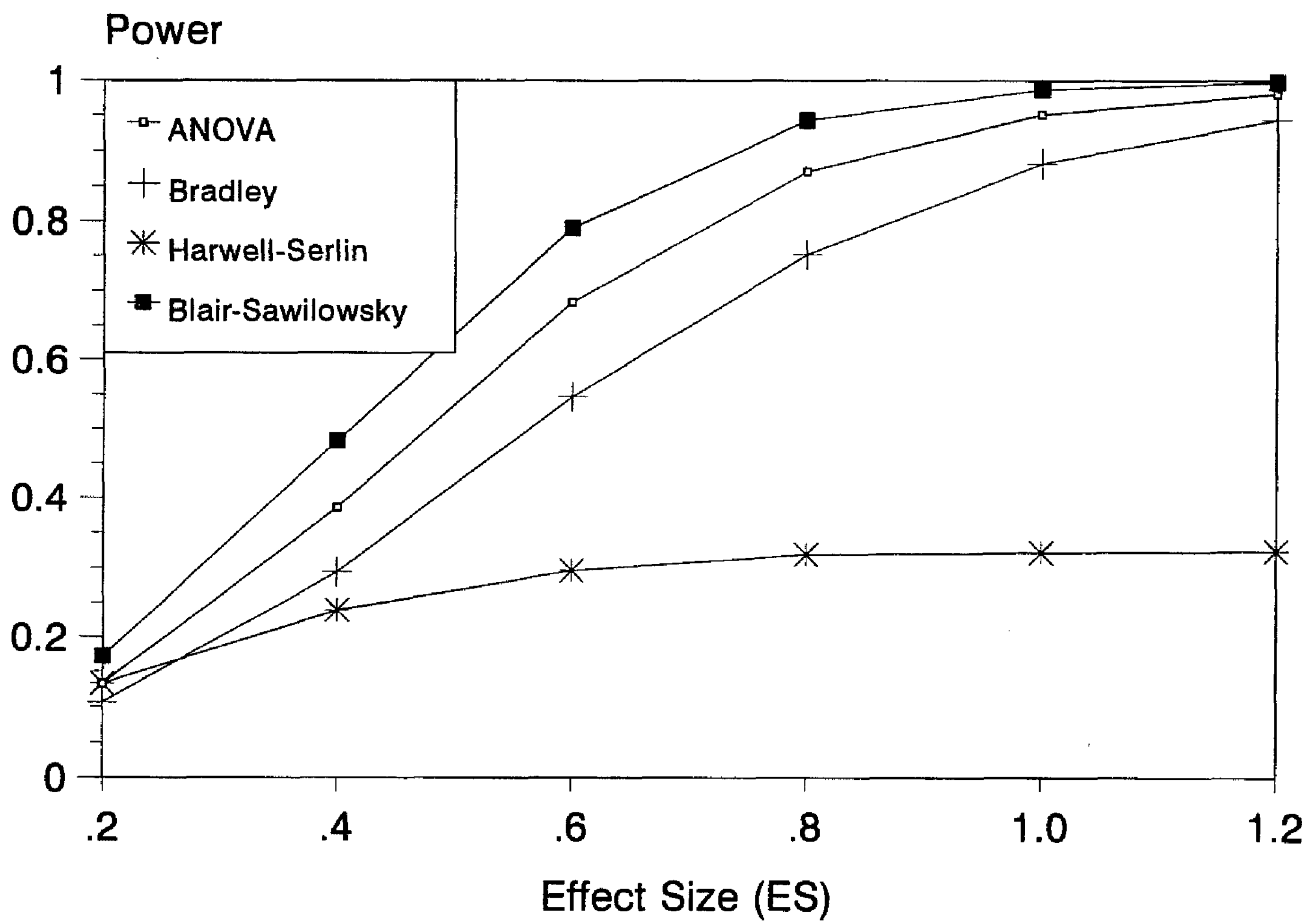

Figure 49. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=7.

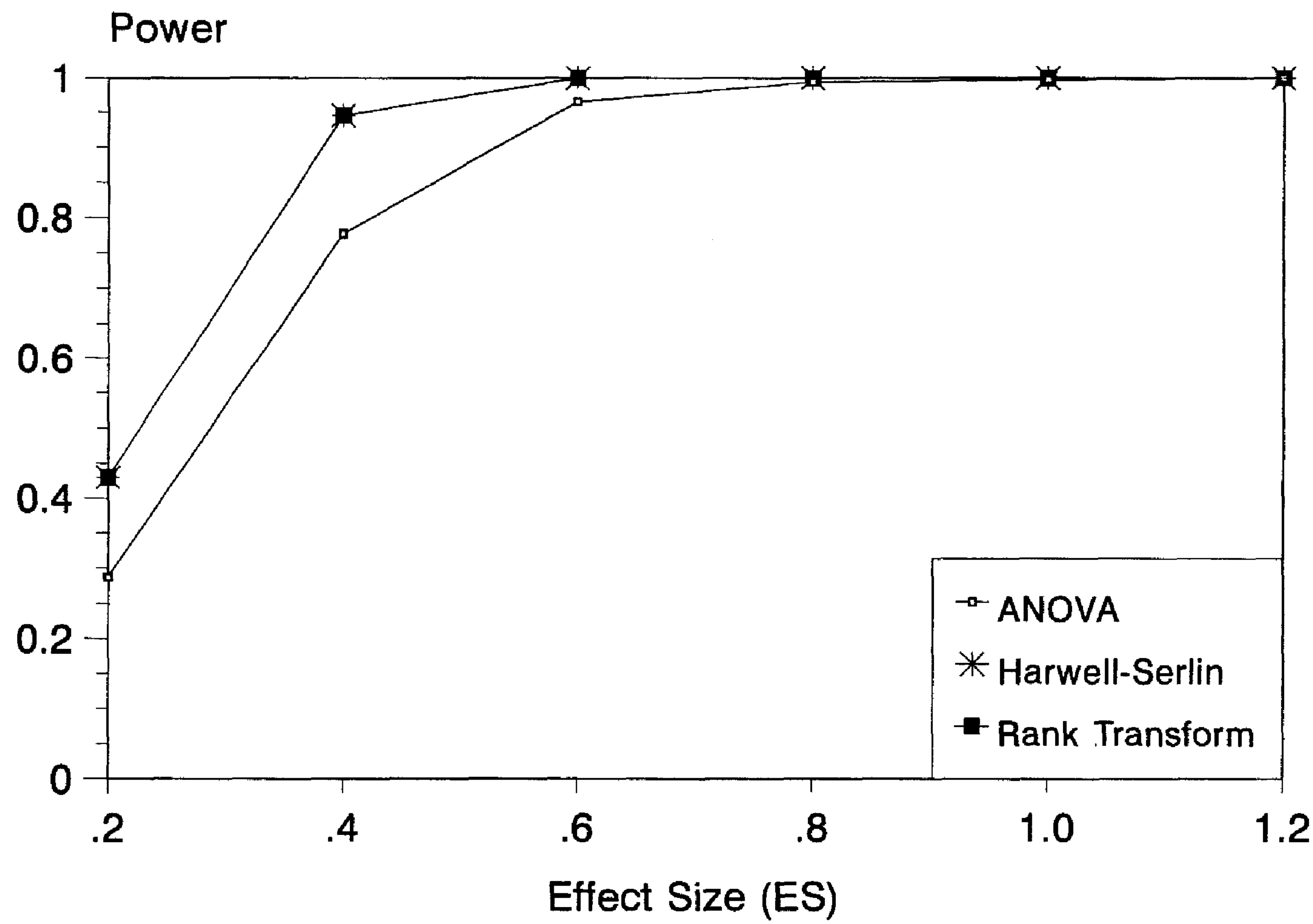

Figure 50. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=21.

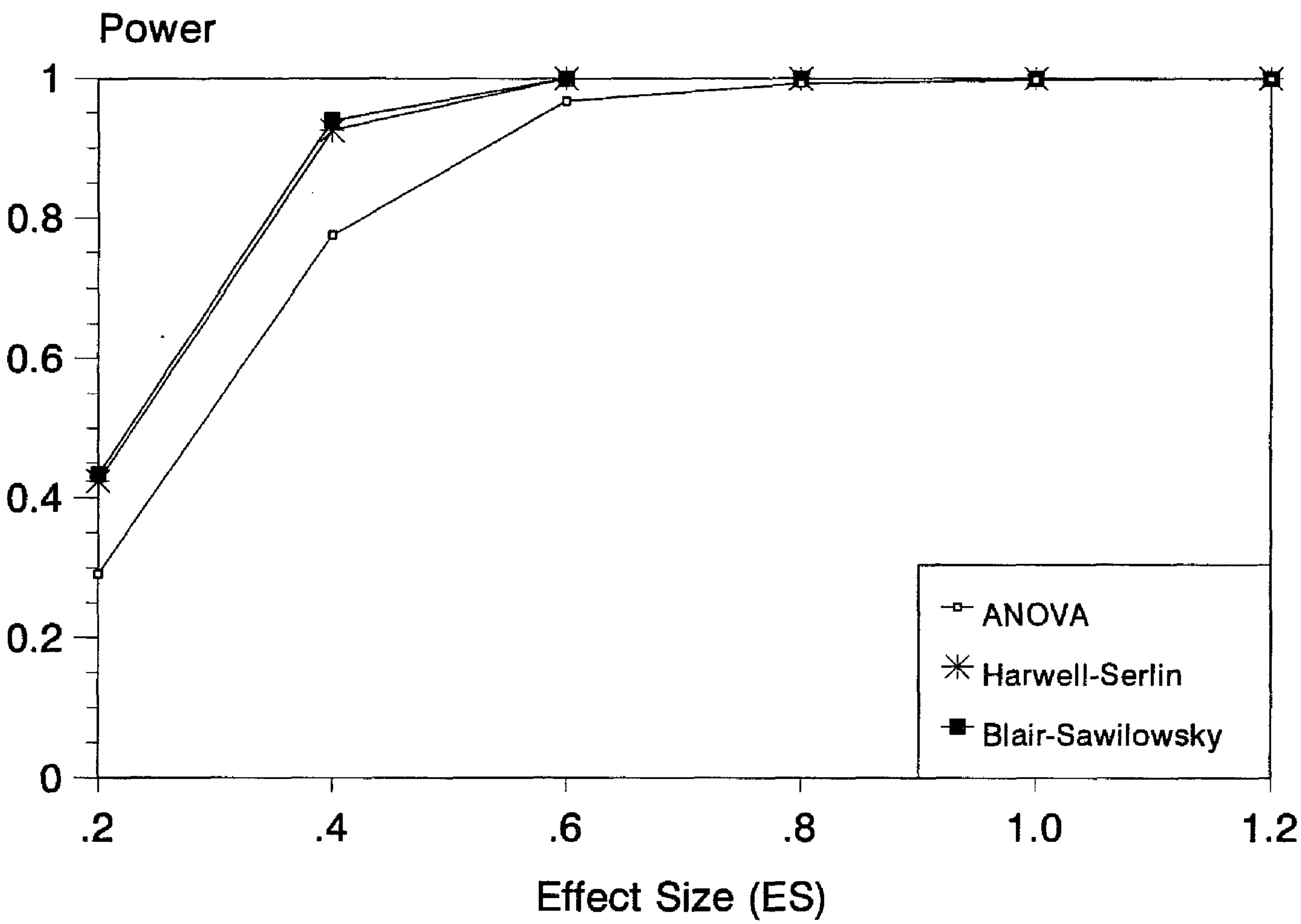

Figure 51. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=21.

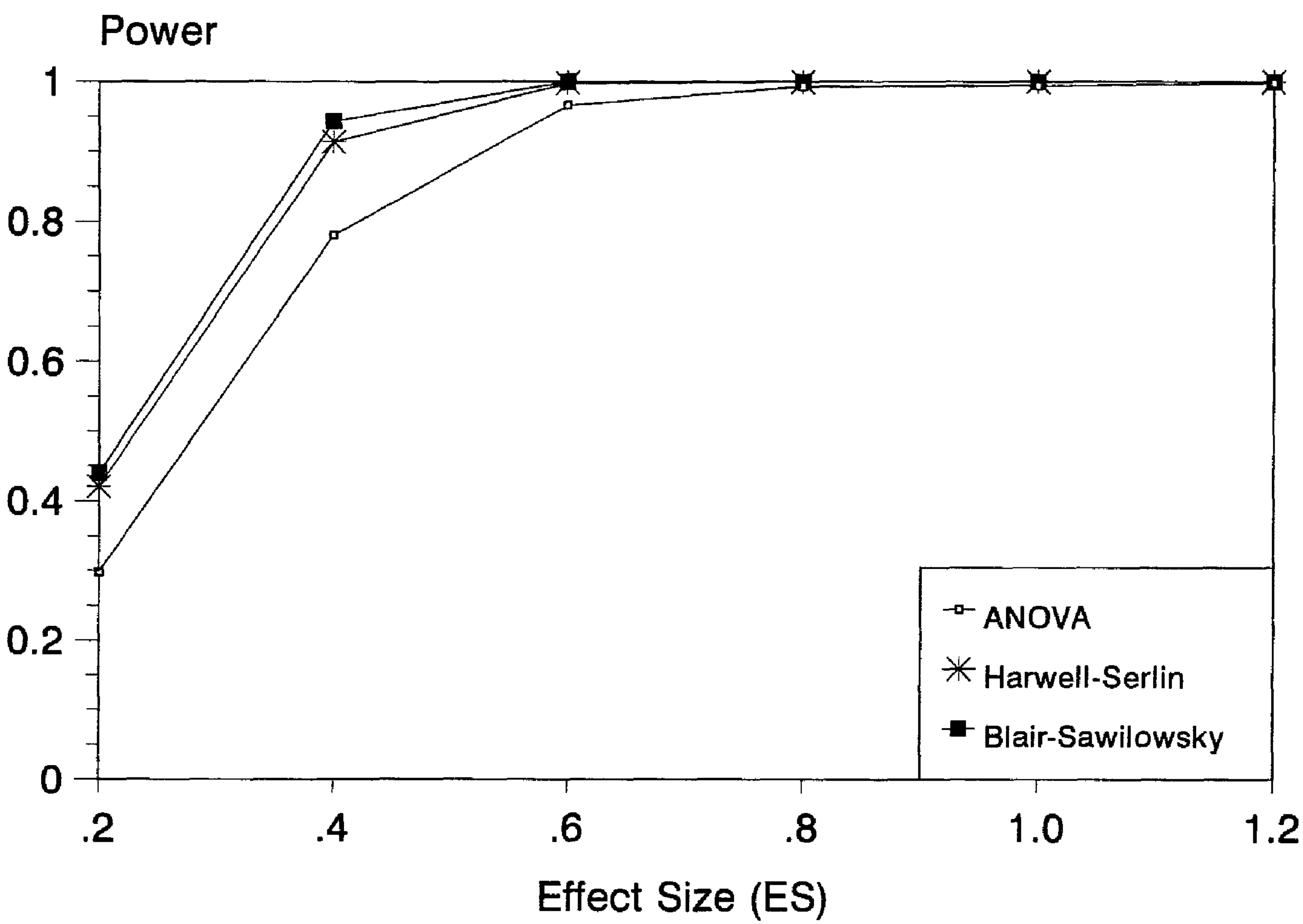

Figure 52. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=21.

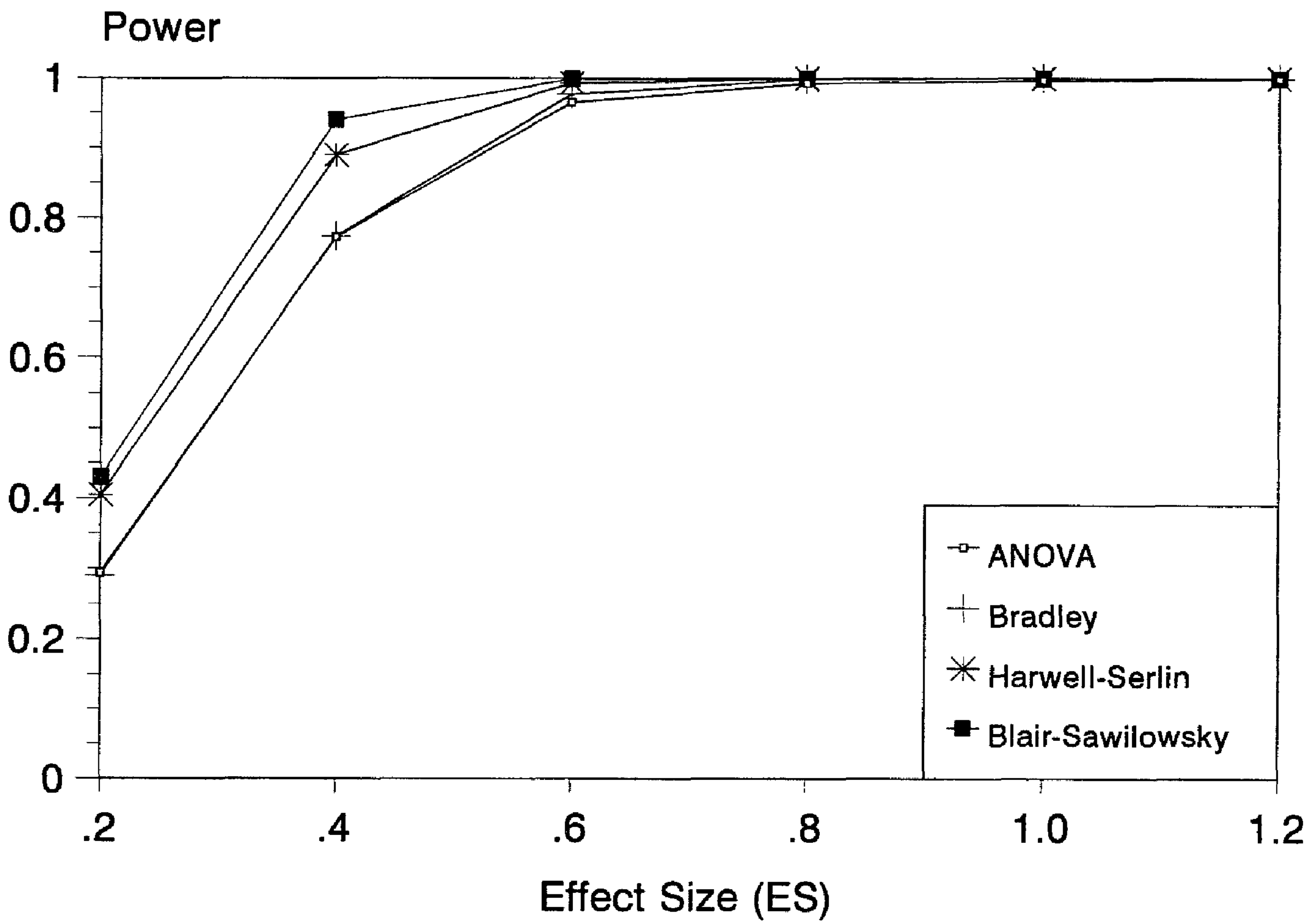

Figure 53. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three dgrees of freedom, alpha=.05 and n=21.

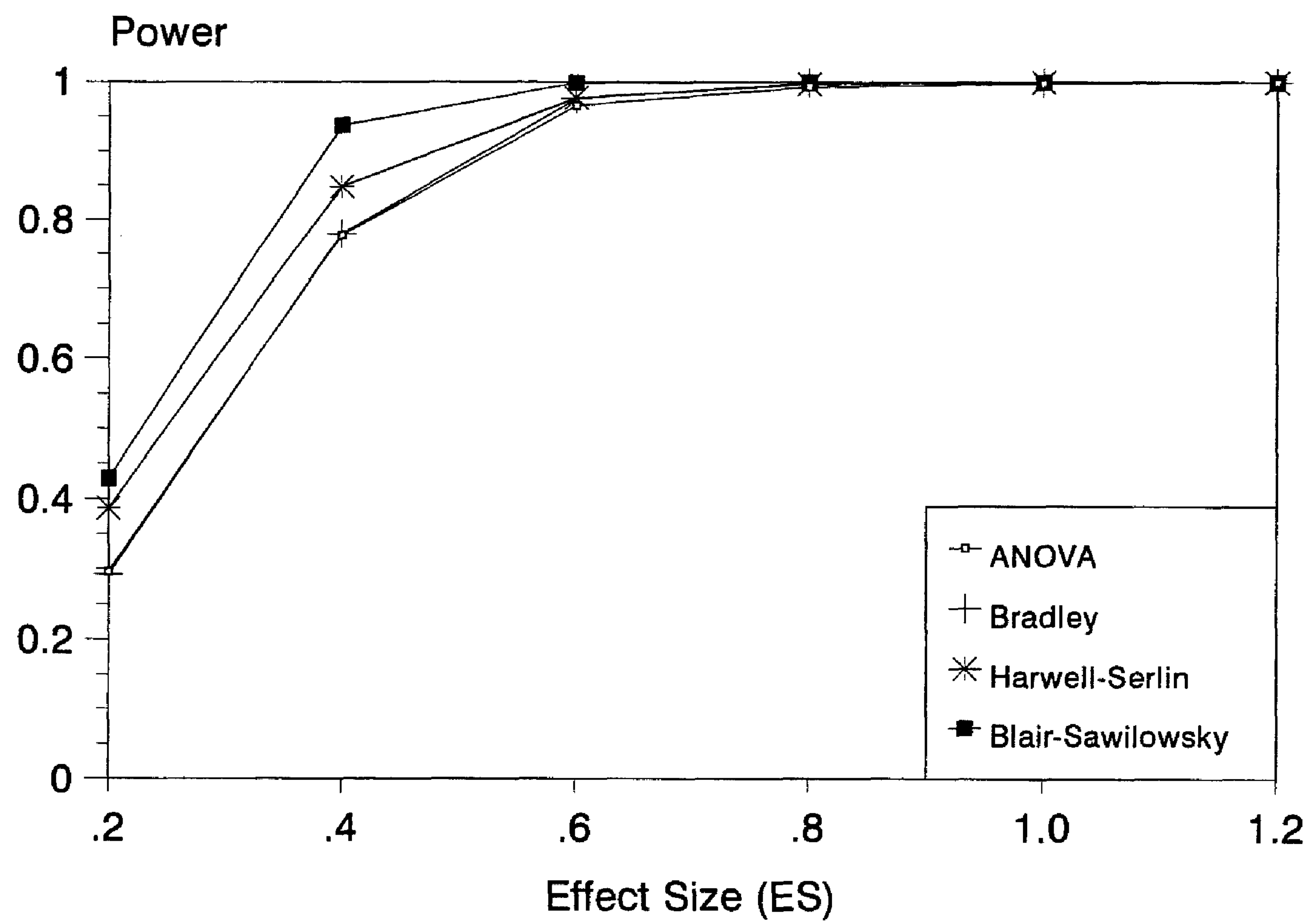

Figure 54. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=21.

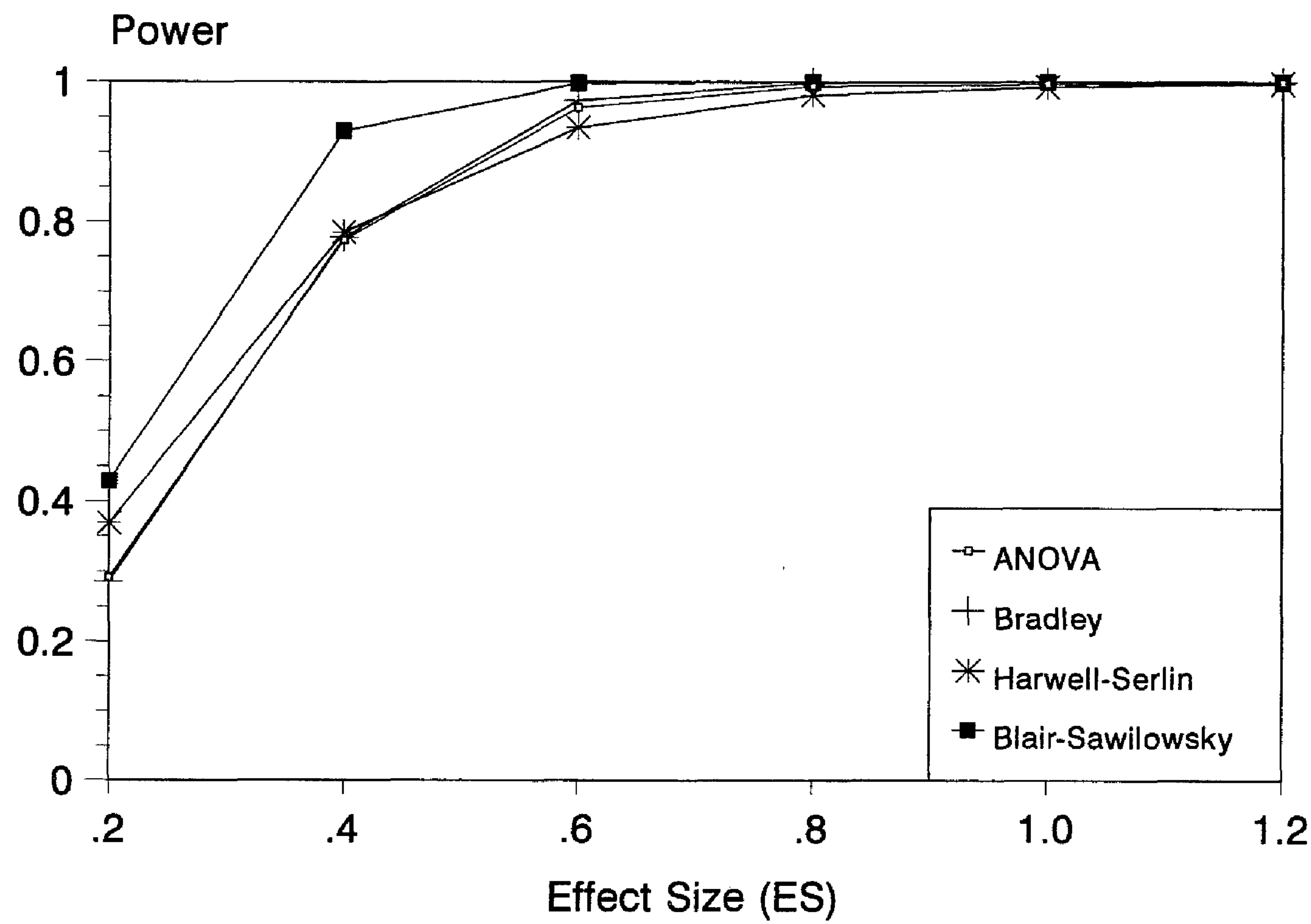

Figure 55. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=21.

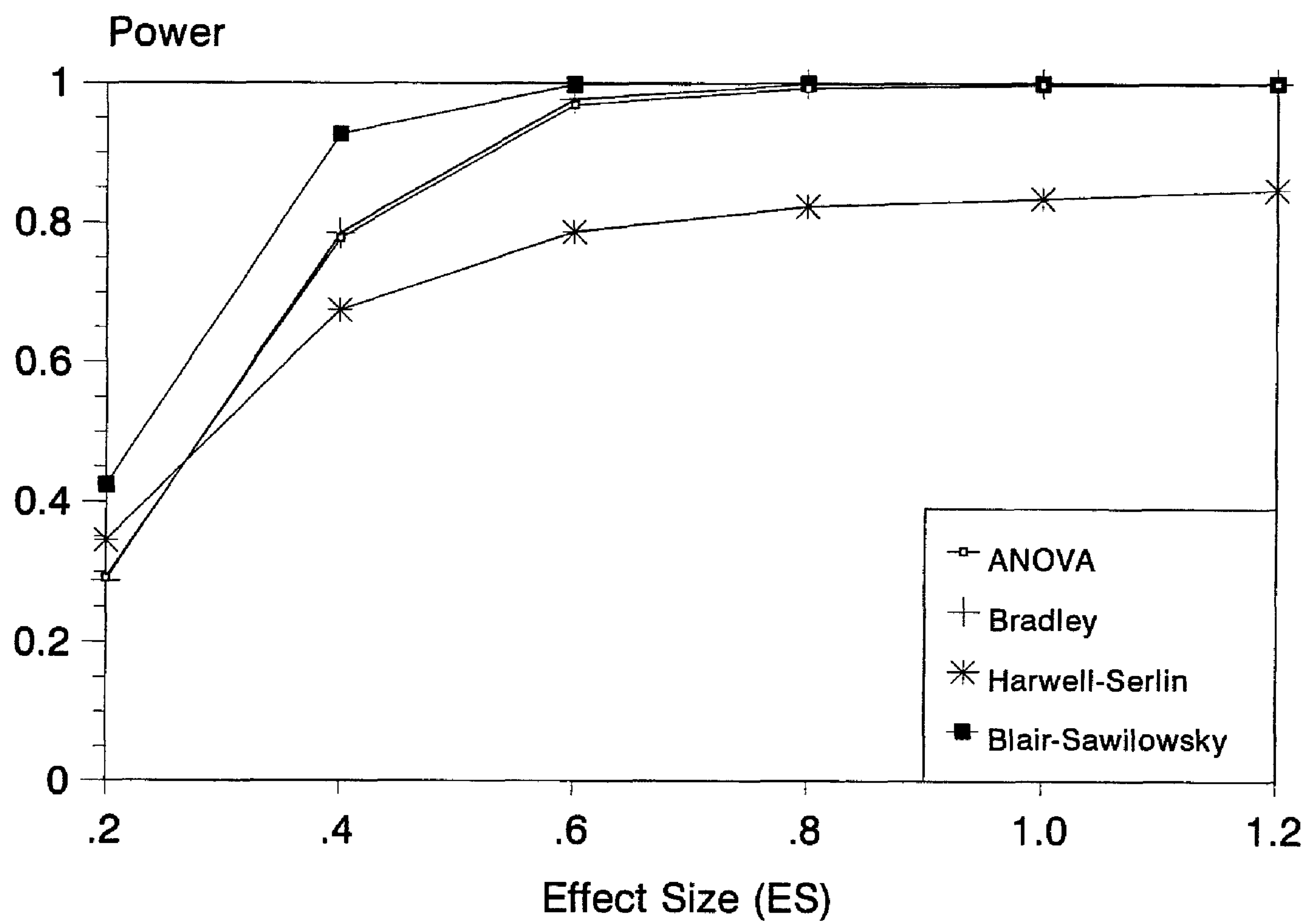

Figure 56. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=7.

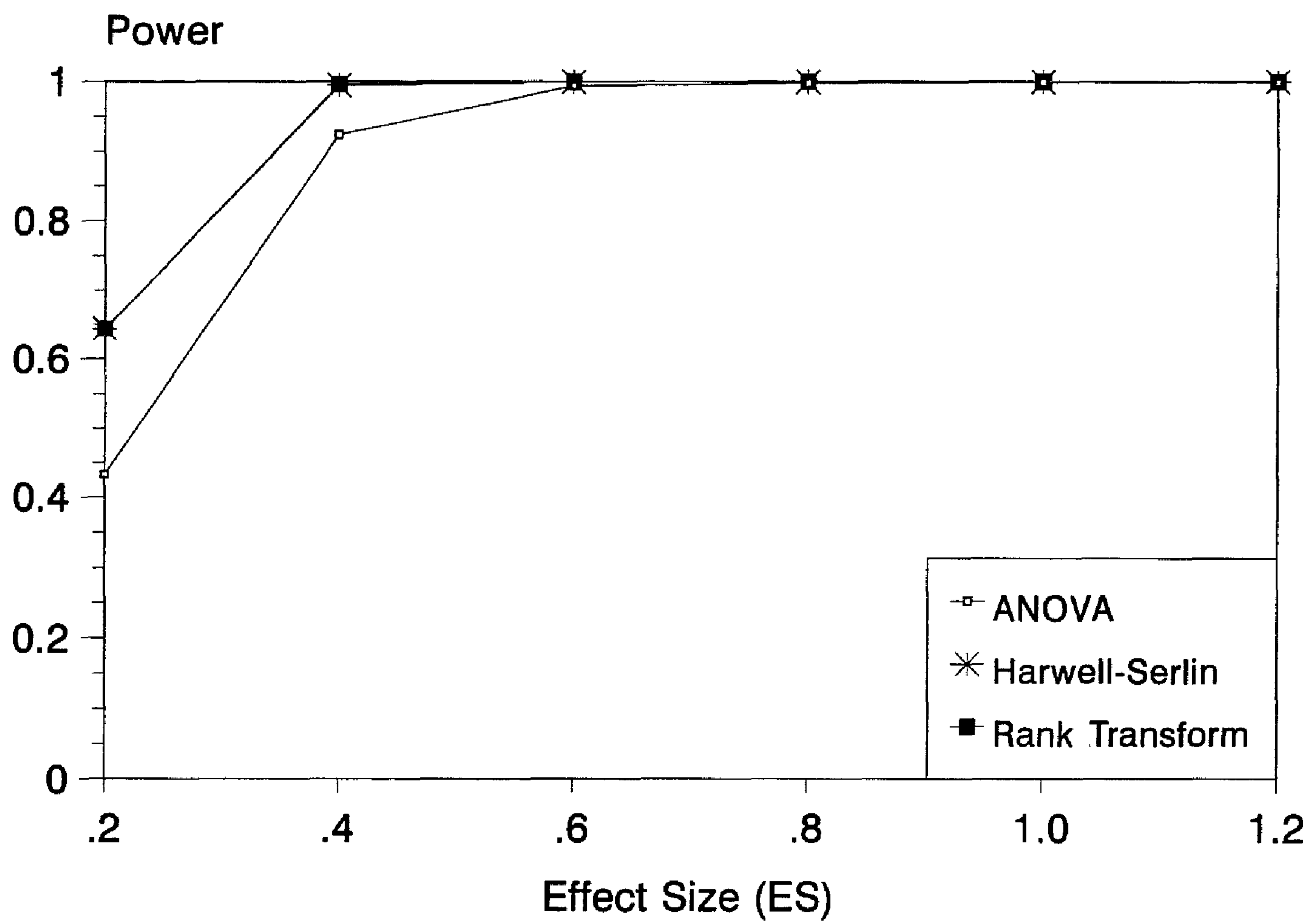

Figure 57. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=35.

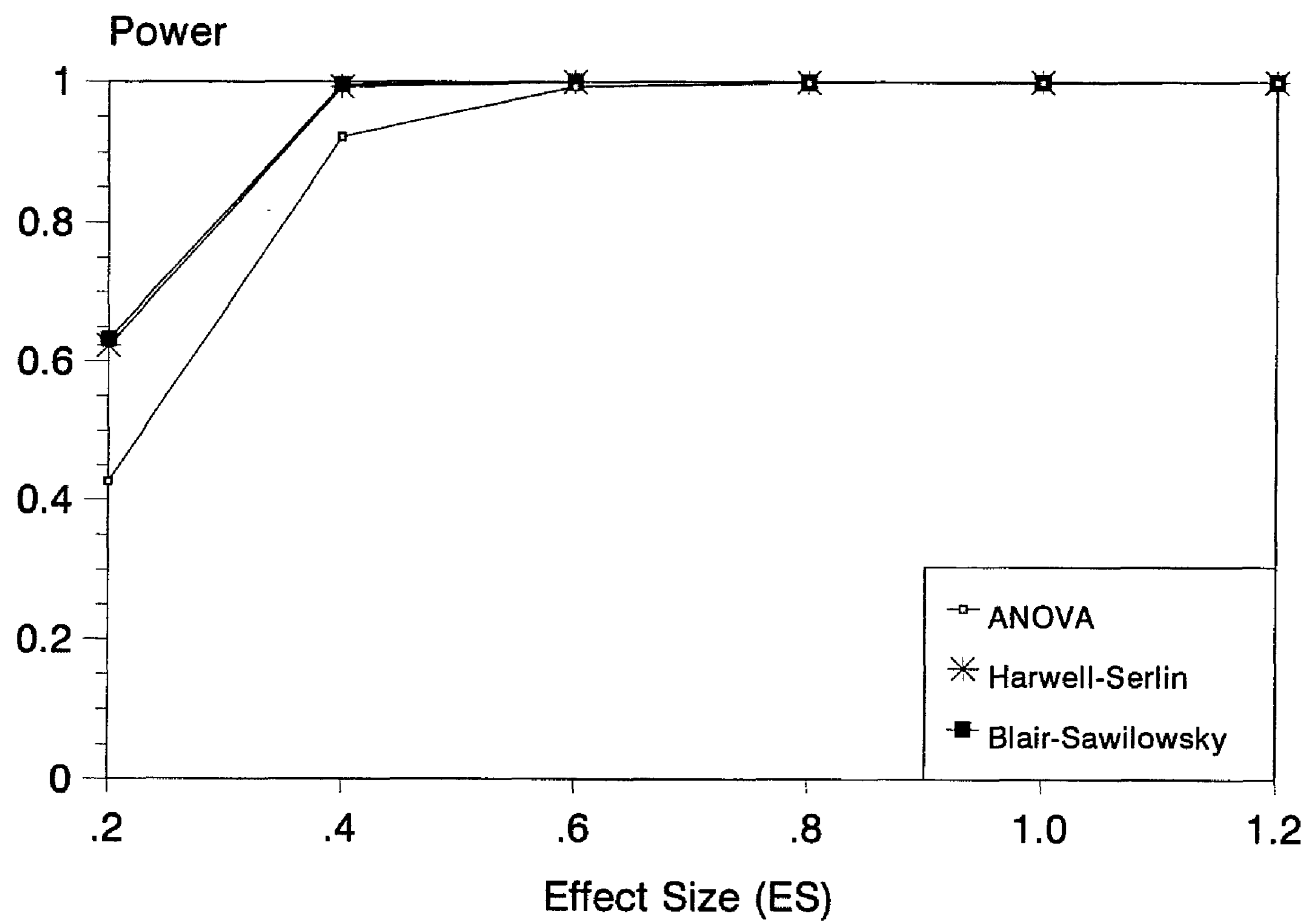

Figure 58. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=35.

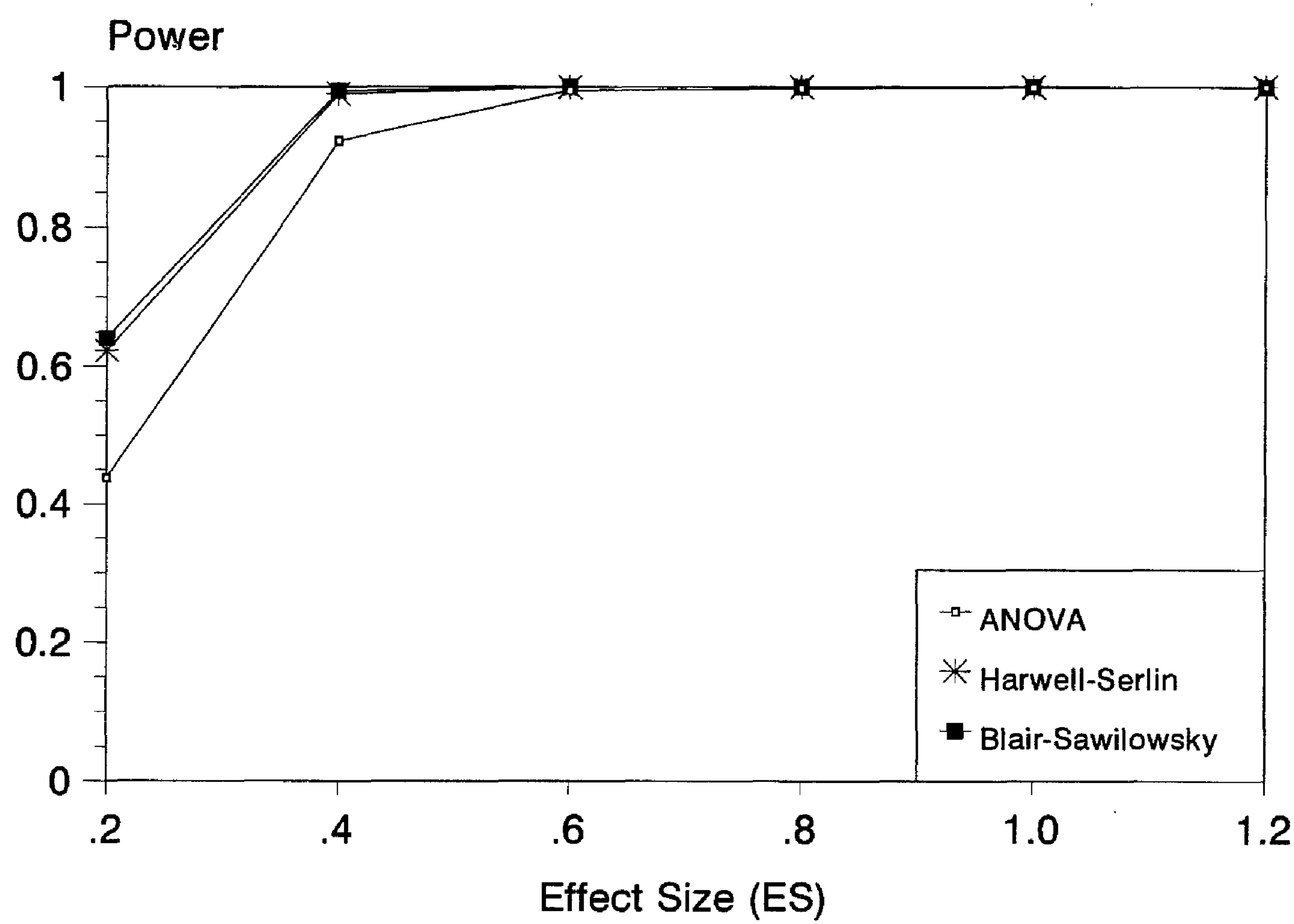

Figure 59. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=35.

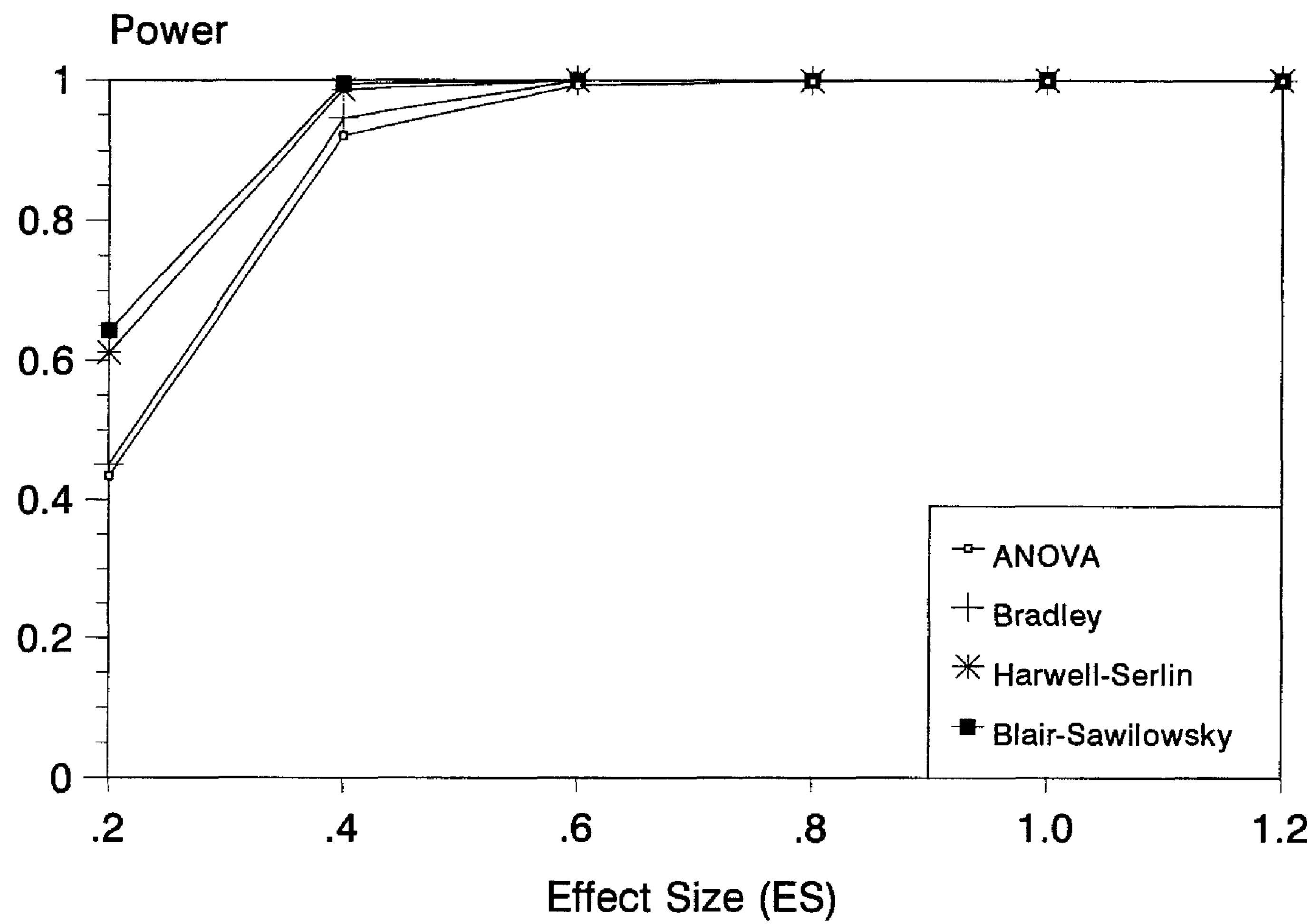

Figure 60. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=35.

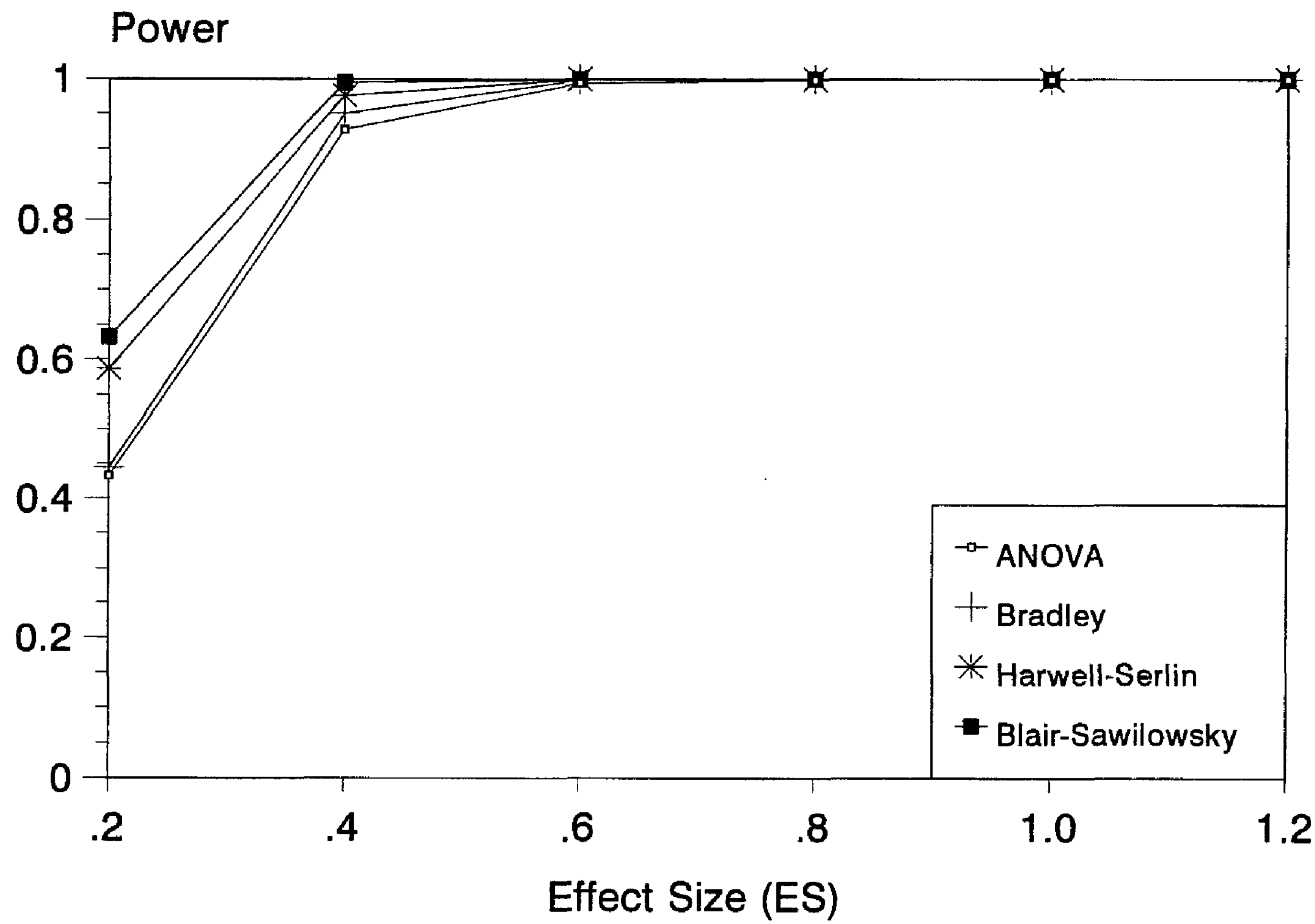

Figure 61. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=35.

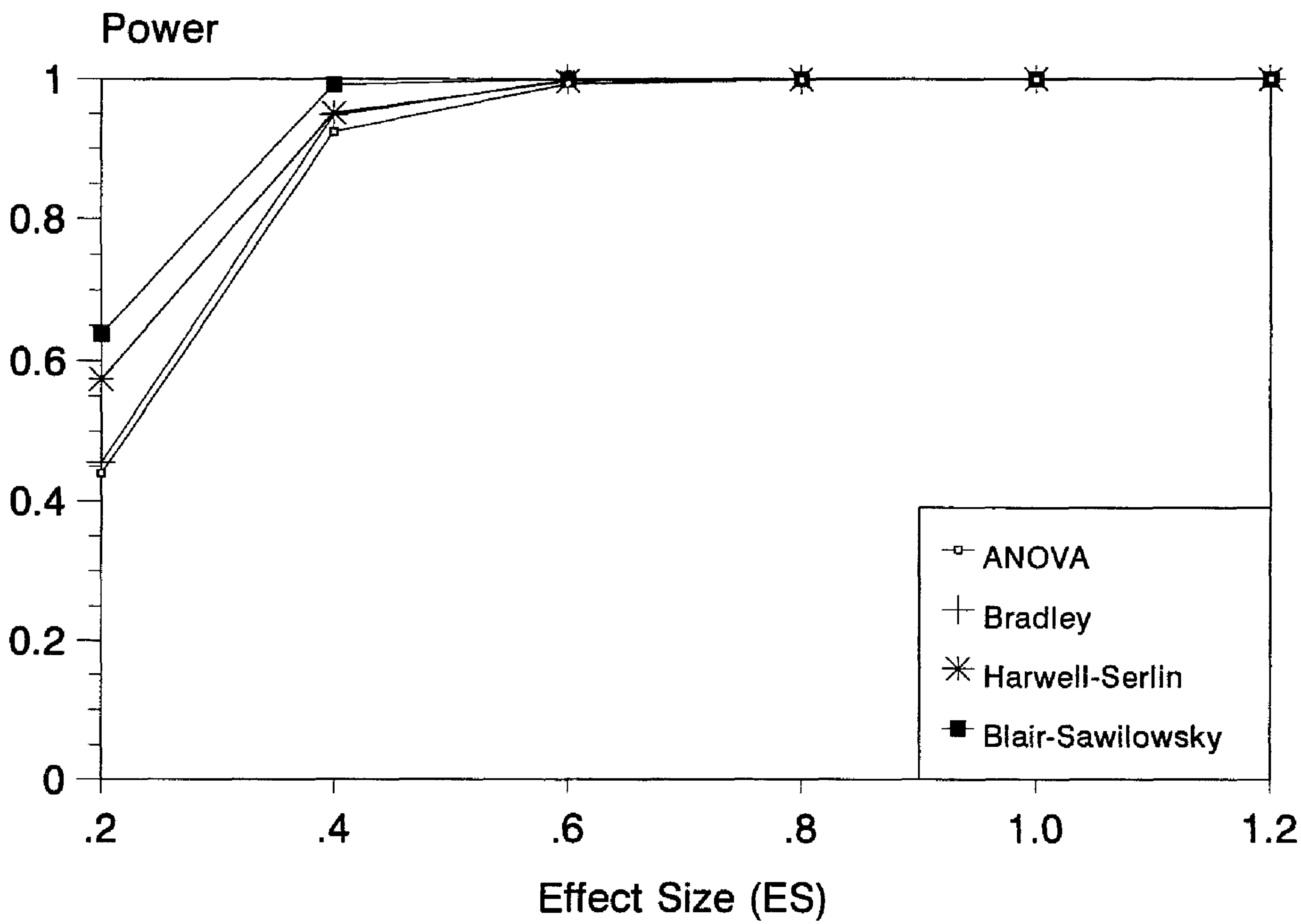

Figure 62. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=35.

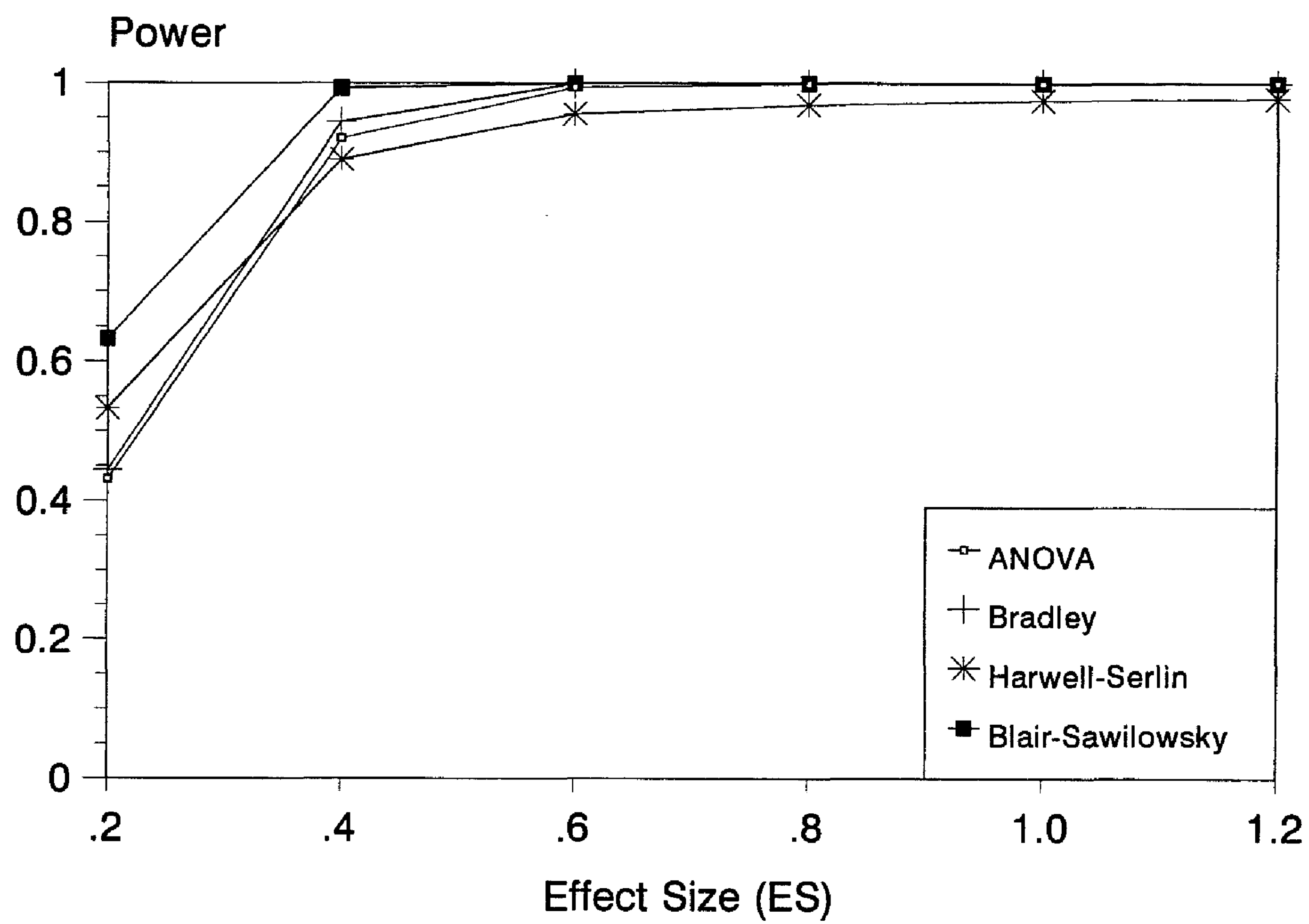

Figure 63. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the t distribution with three degrees of freedom, alpha=.05 and n=35.

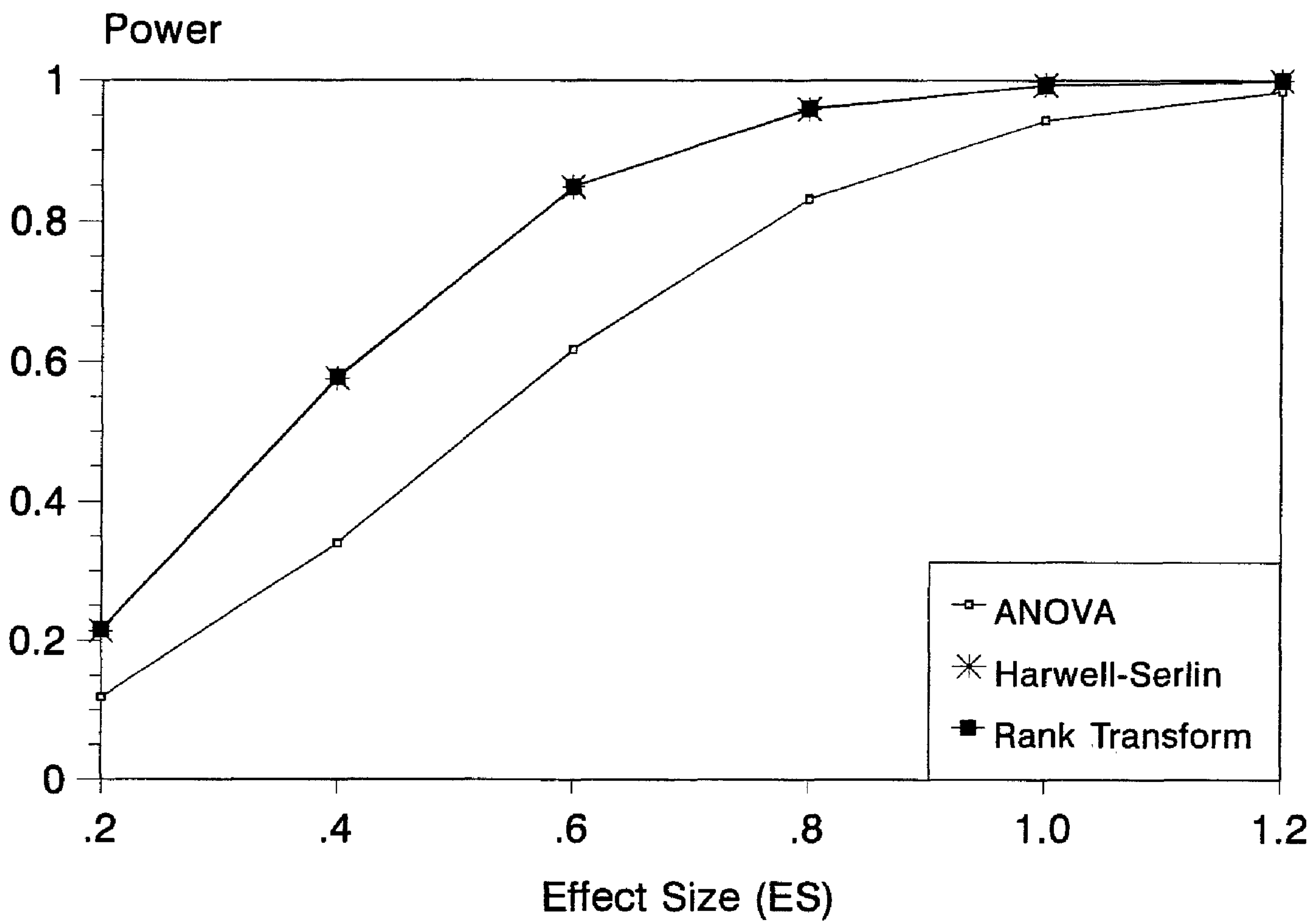

Figure 64. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the exponential distribution, alpha=.05 and n=7.

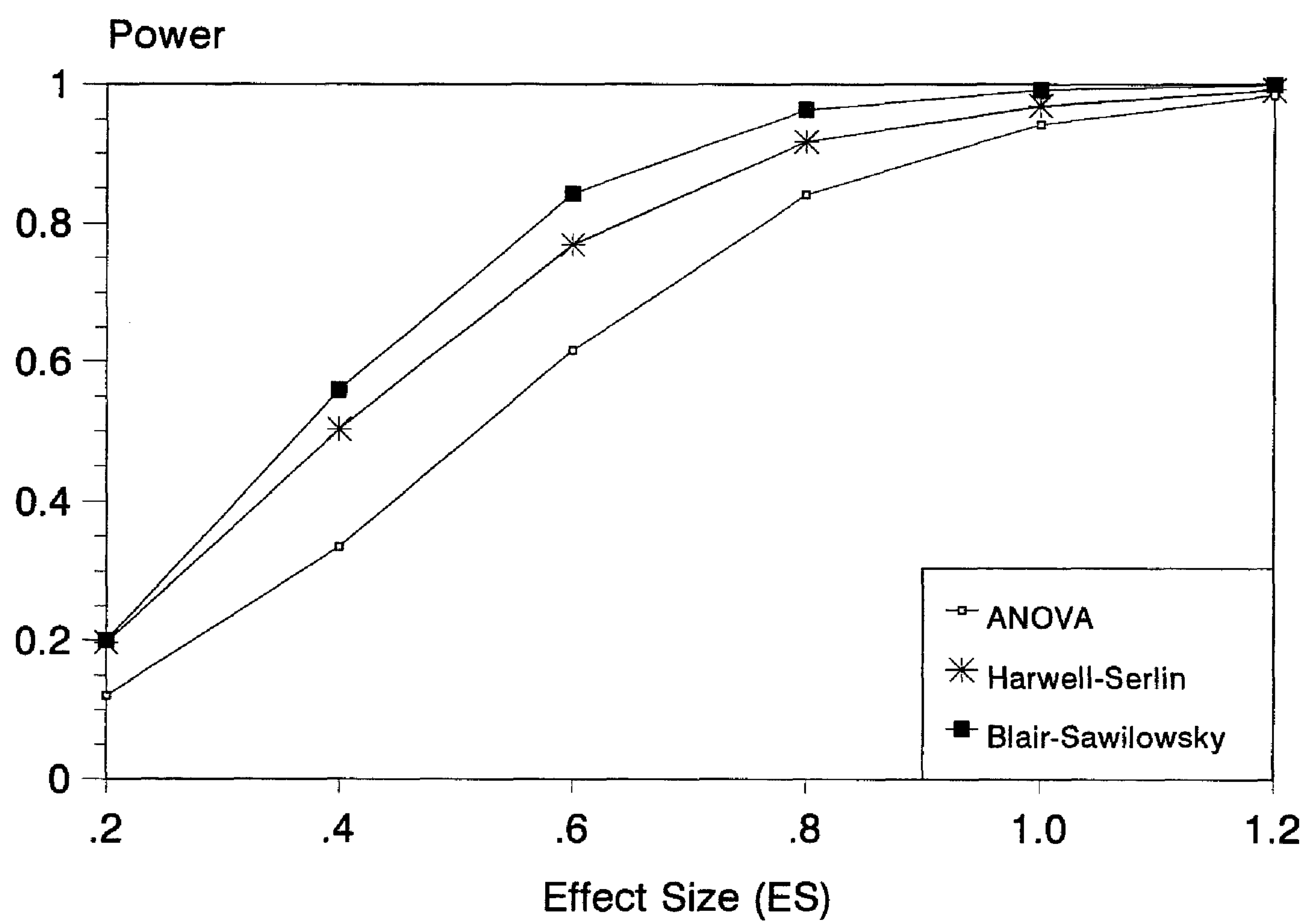

Figure 65. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=7.

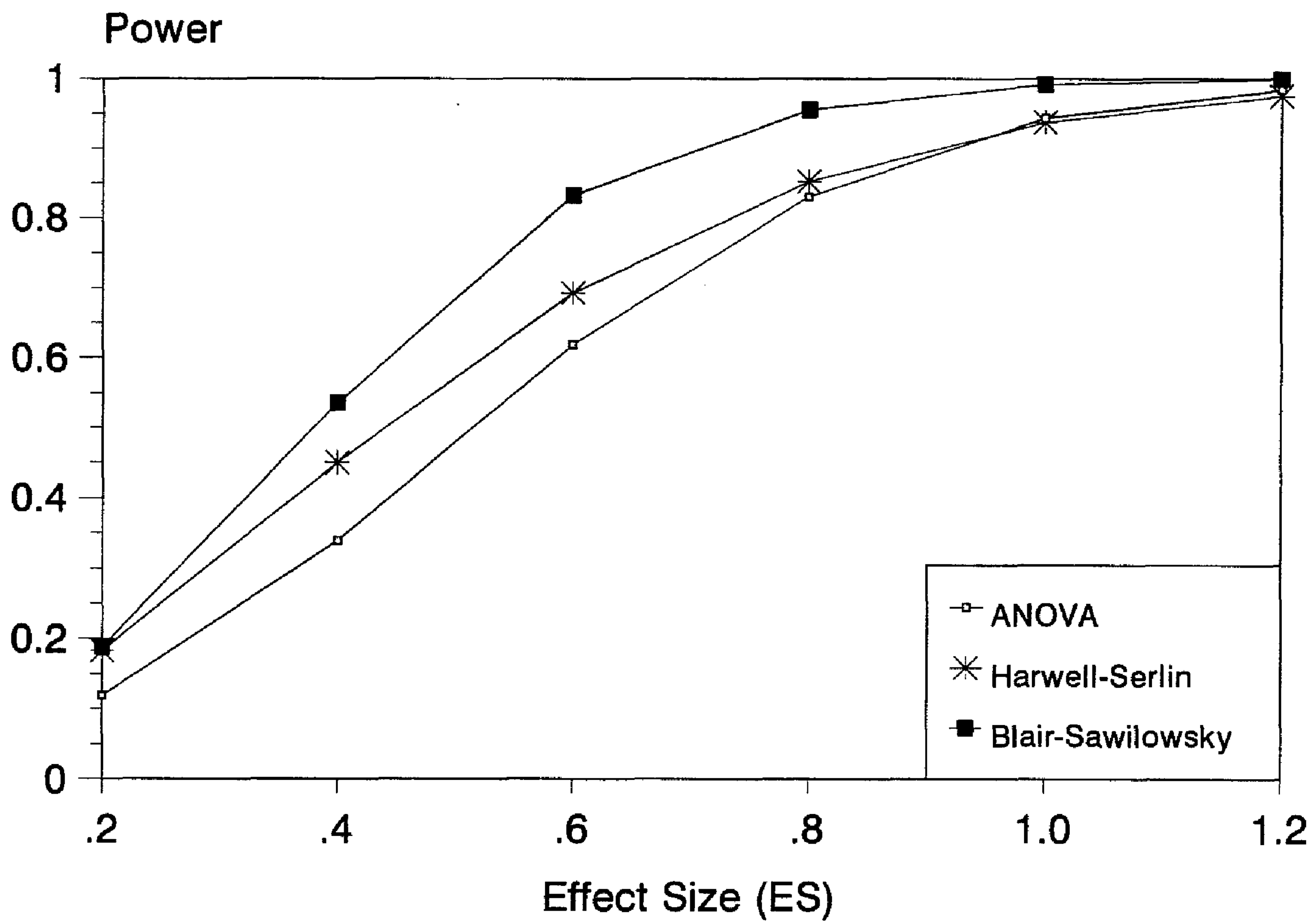

Figure 66. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=7.

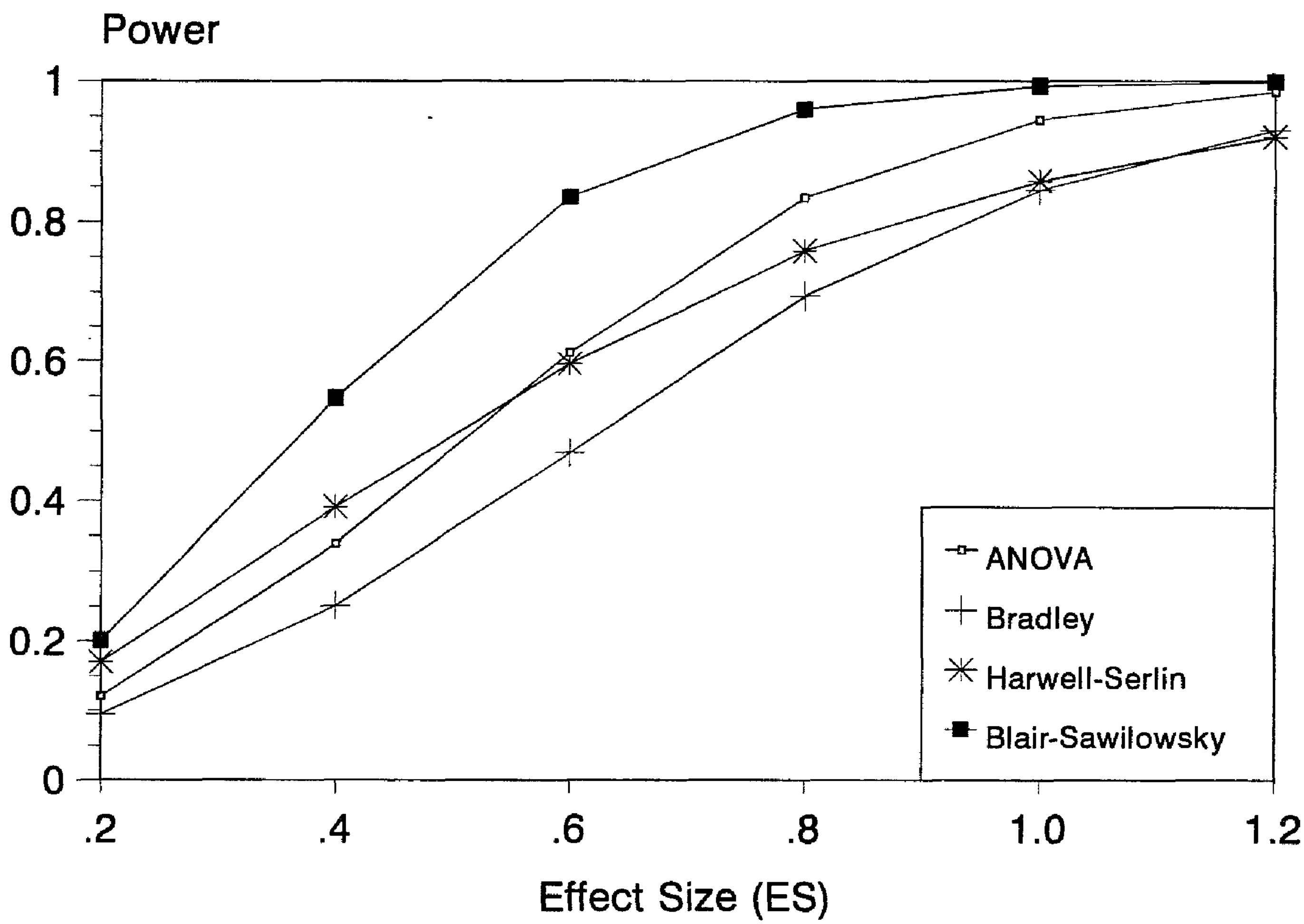

Figure 67. Comparative power of the (bc) interation for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=7.

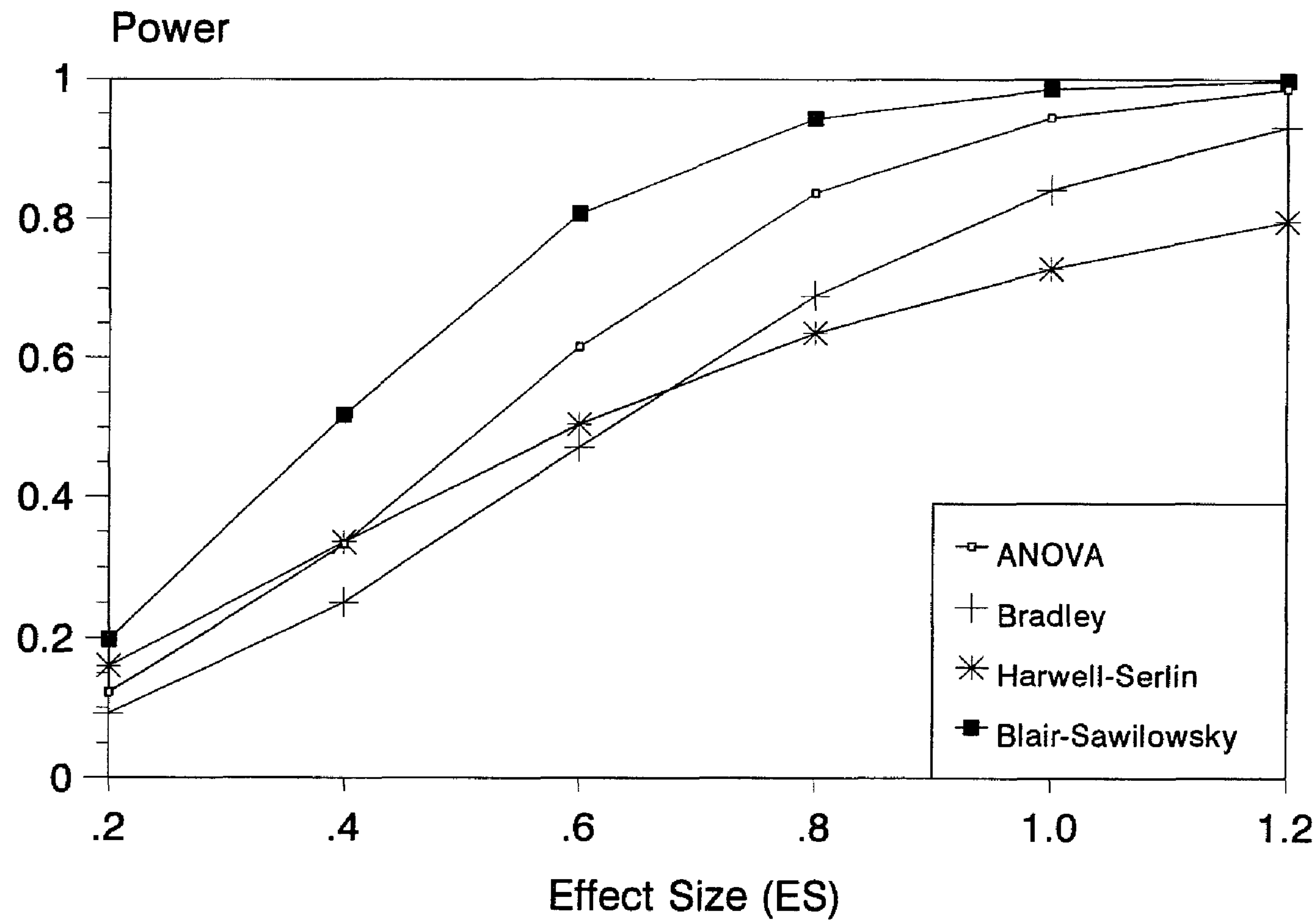

Figure 68. Comparative power of the (ac) interation for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=7.

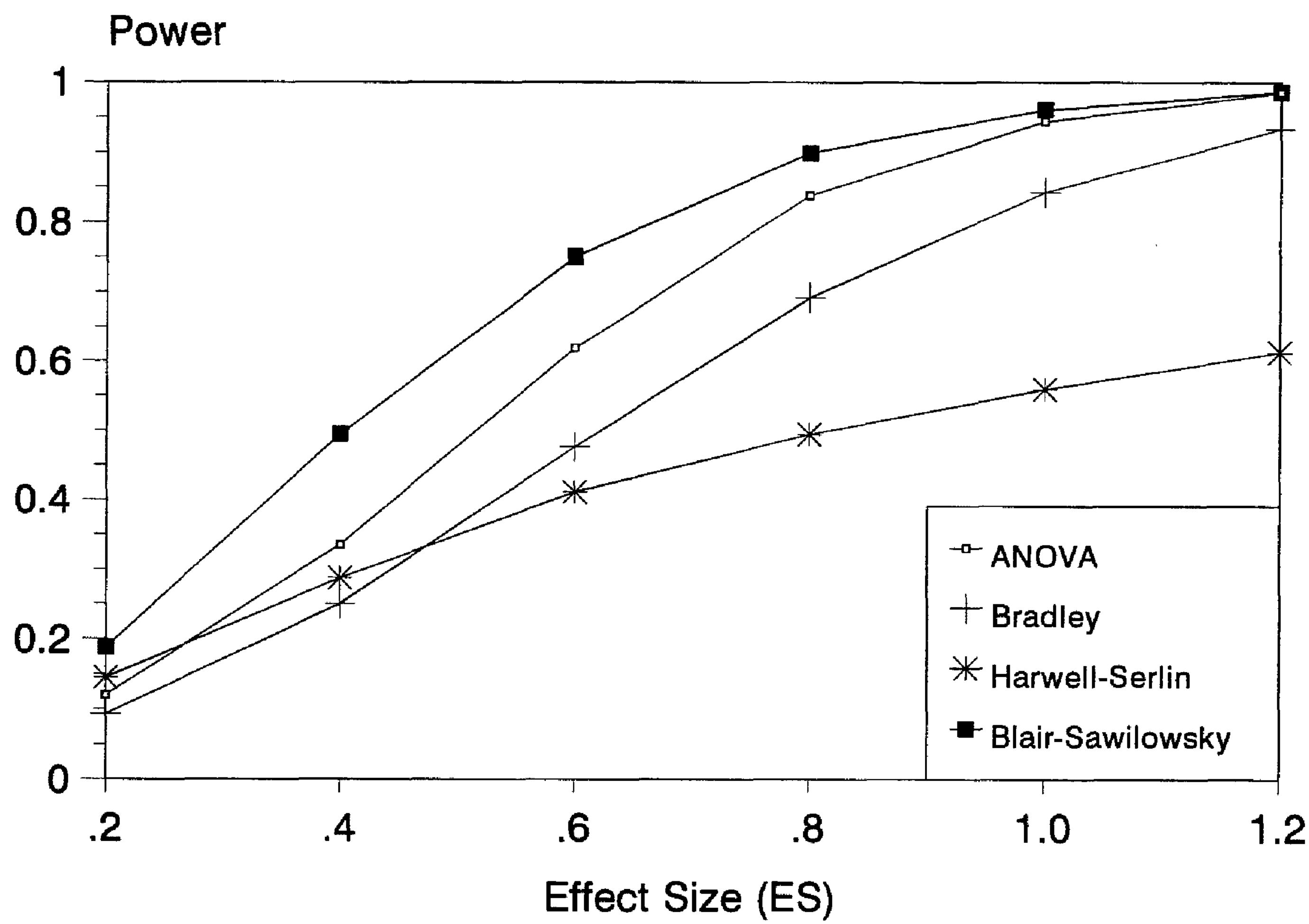

Figure 69. Comparative power of the (ab) interation for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=7.

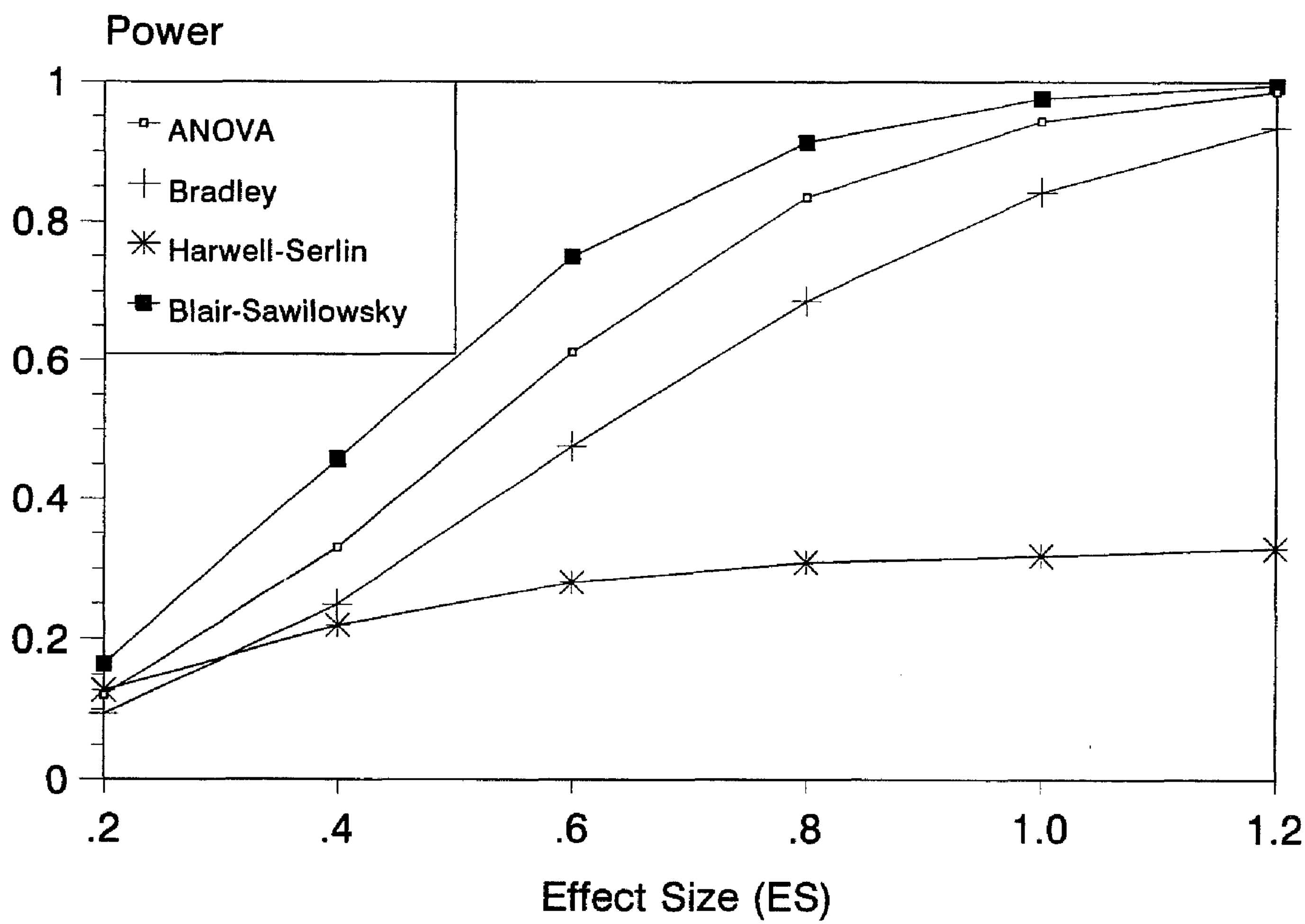

Figure 70. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=7.

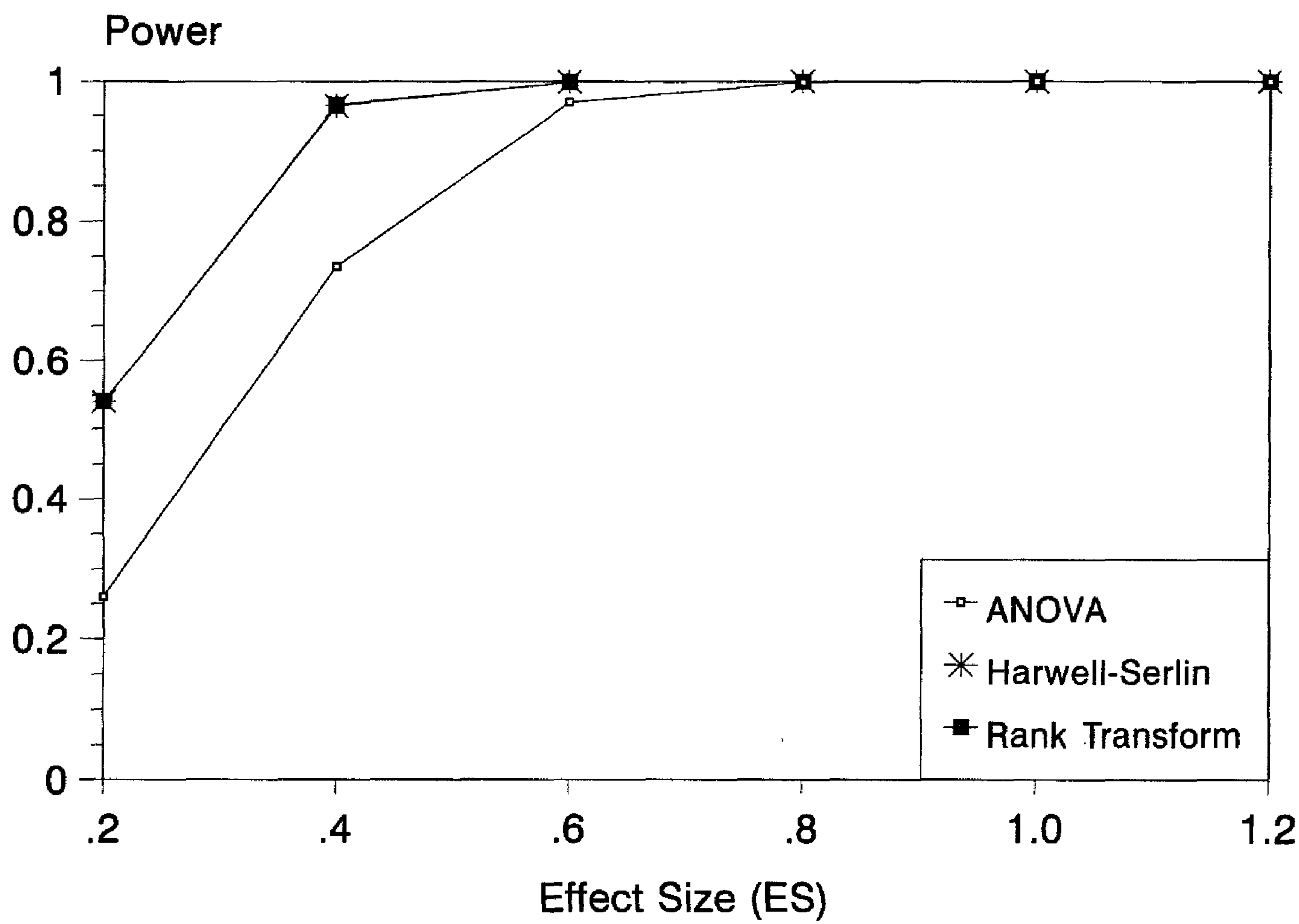

Figure 71. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the exponential distribution, alpha=.05 and n=21.

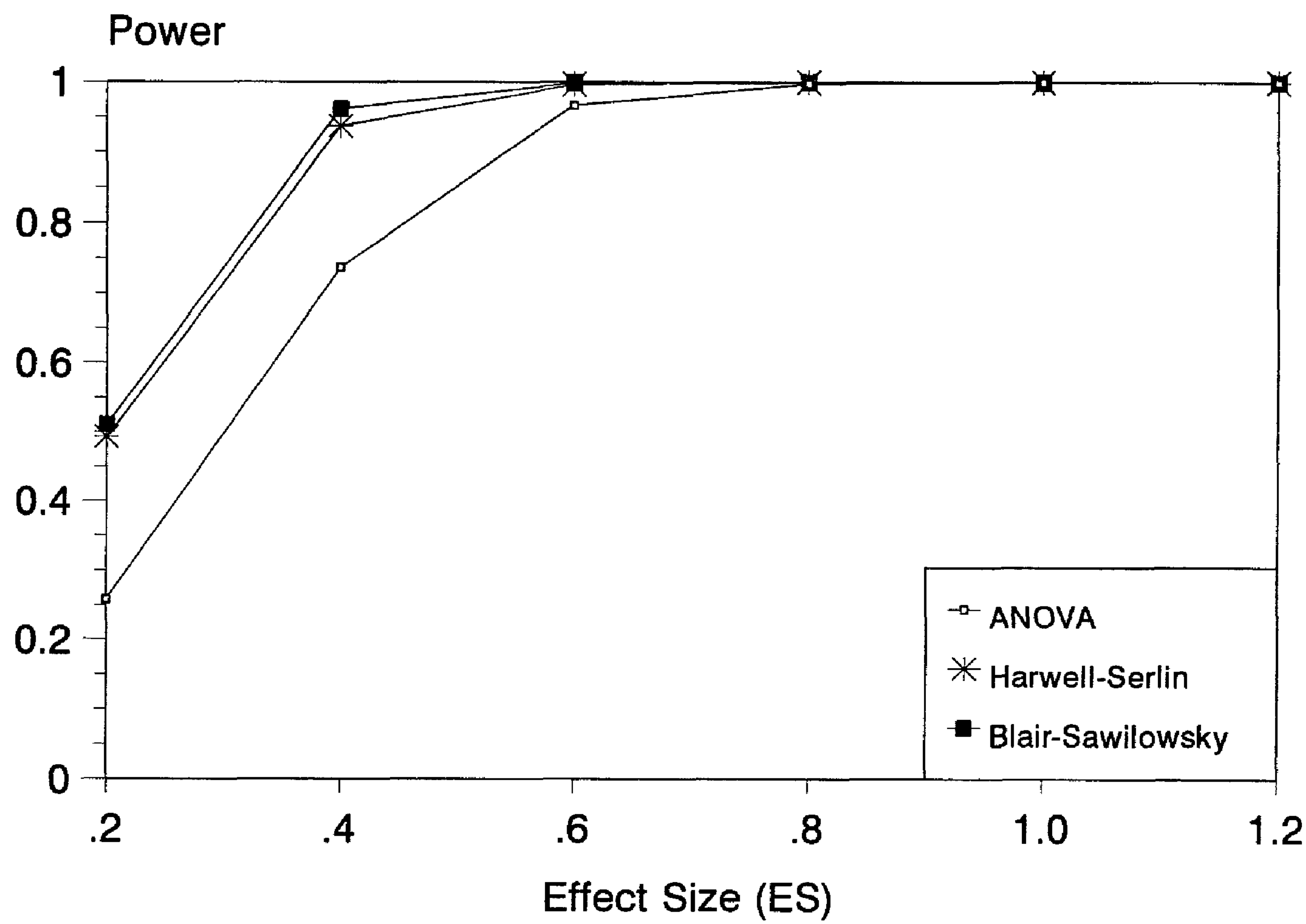

Figure 72. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=21.

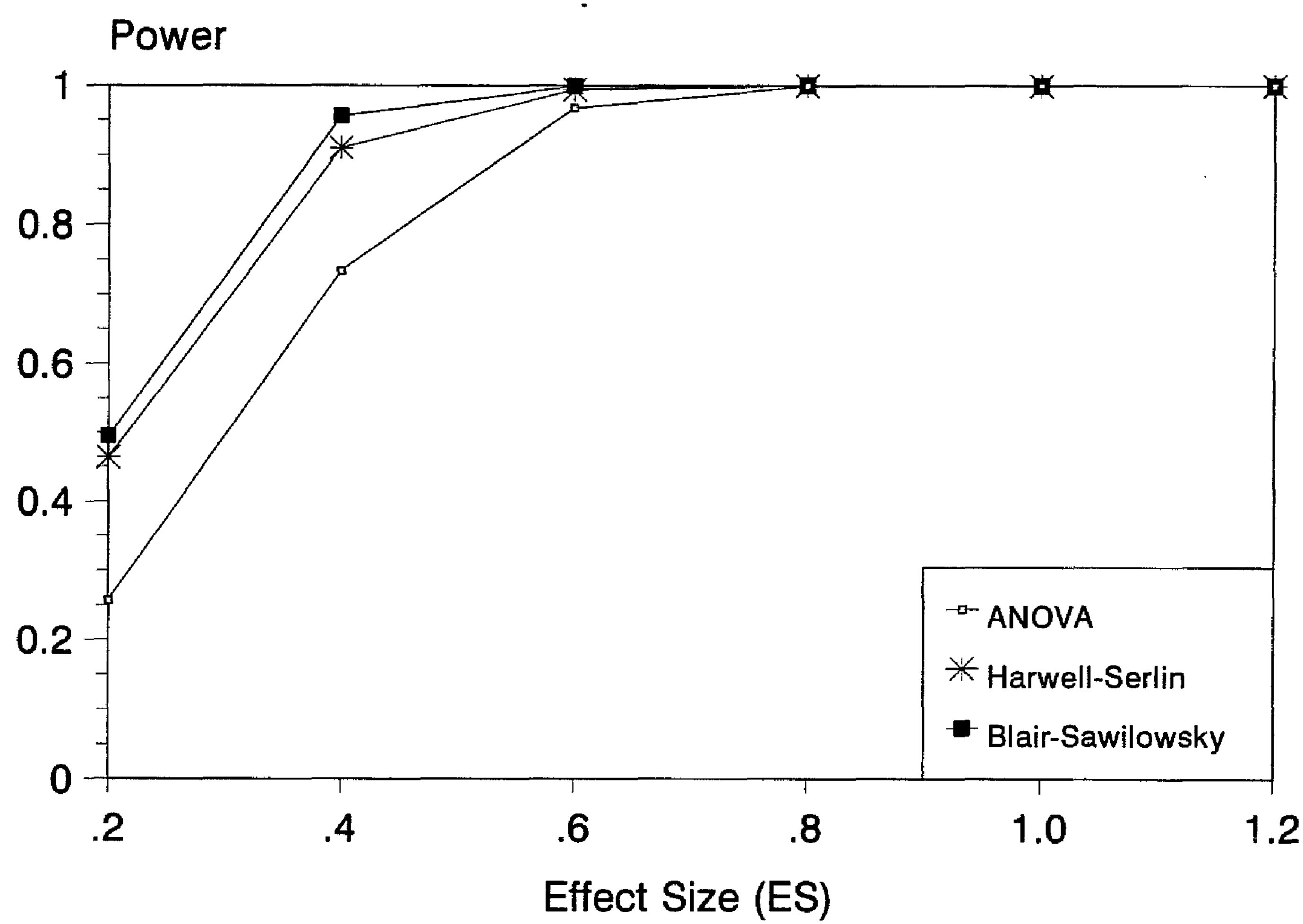

Figure 73. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=21.

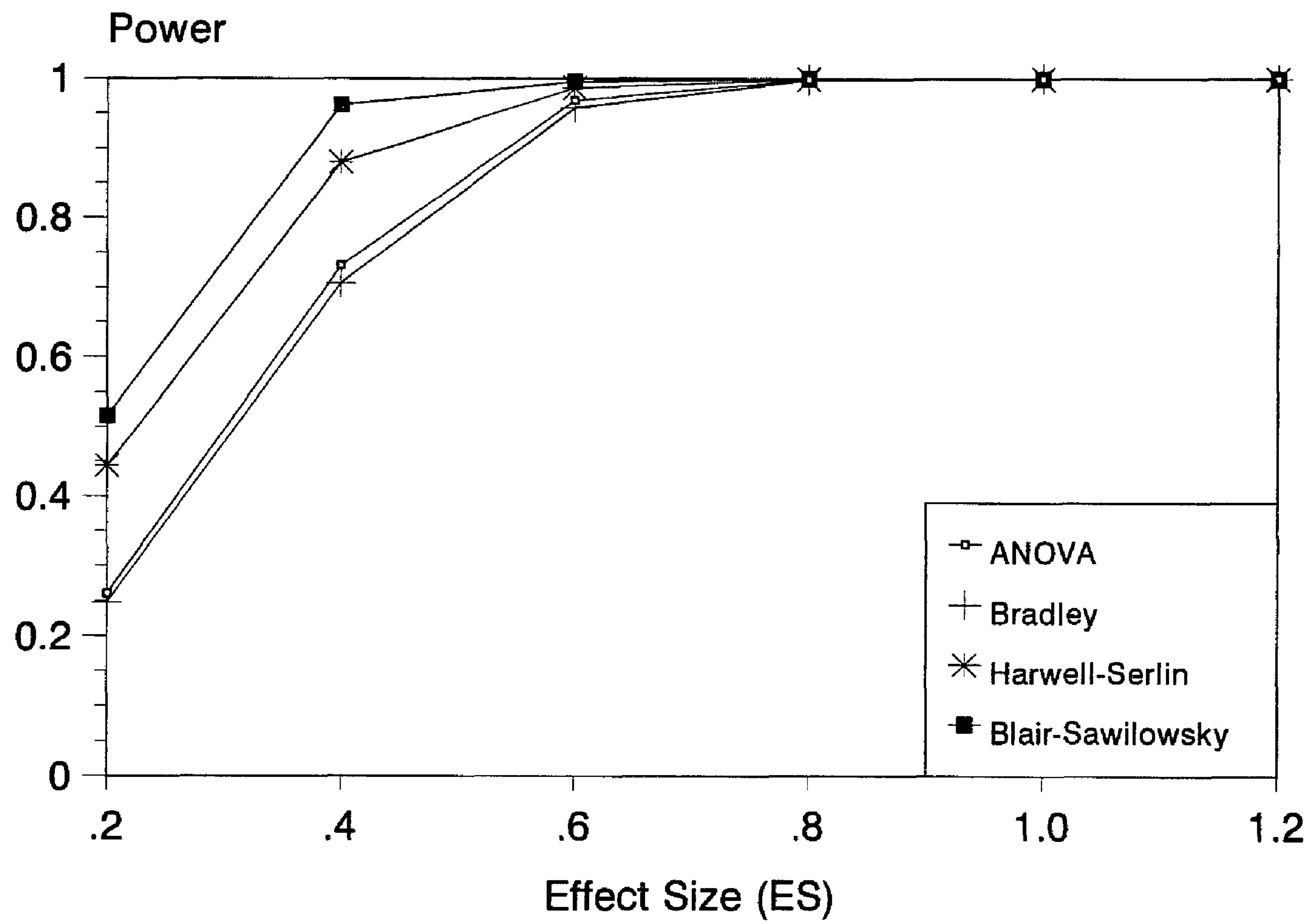

Figure 74. Comparative power of the (bc) interation for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=21.

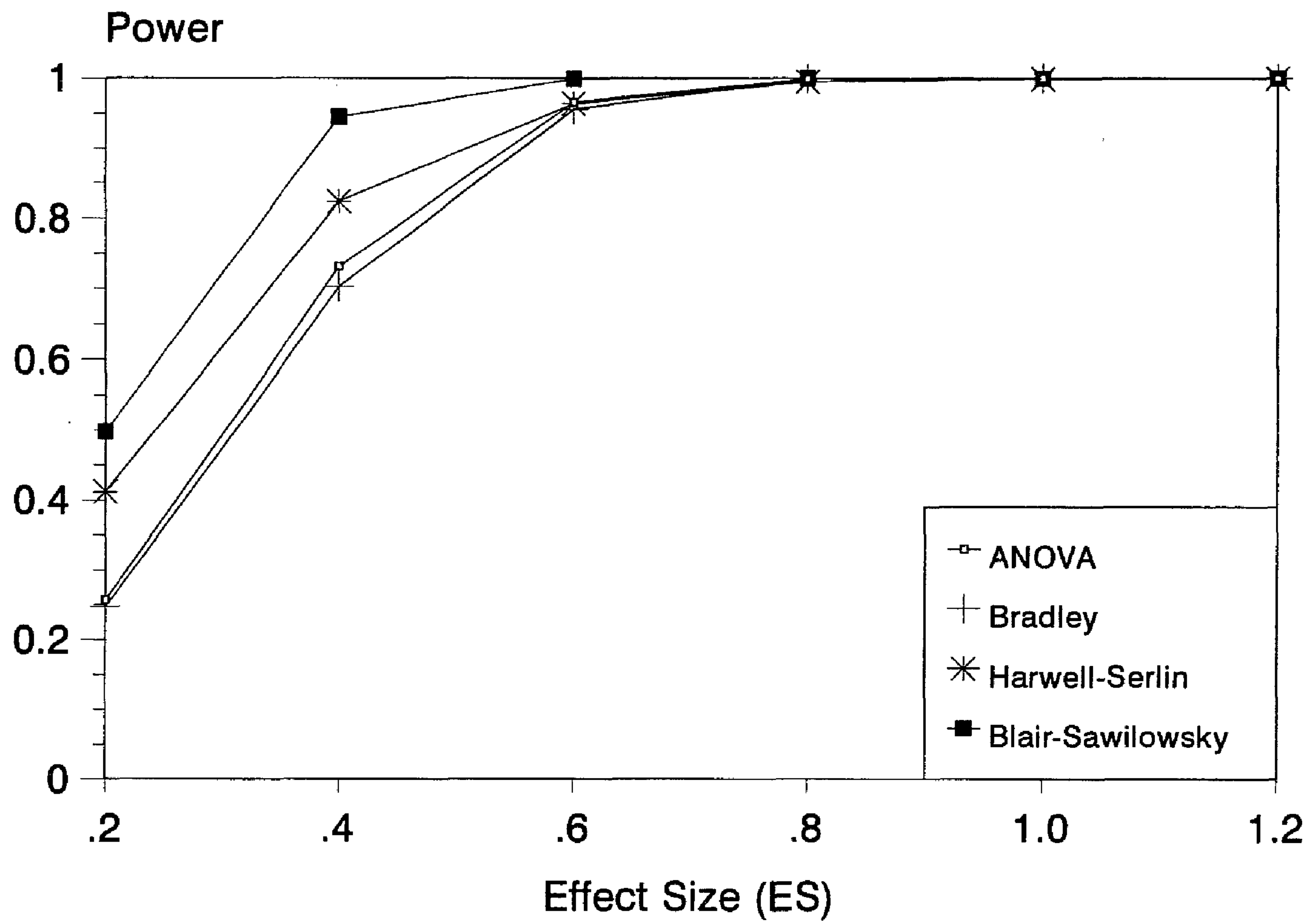

Figure 75. Comparative power of the (ac) interation for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=21.

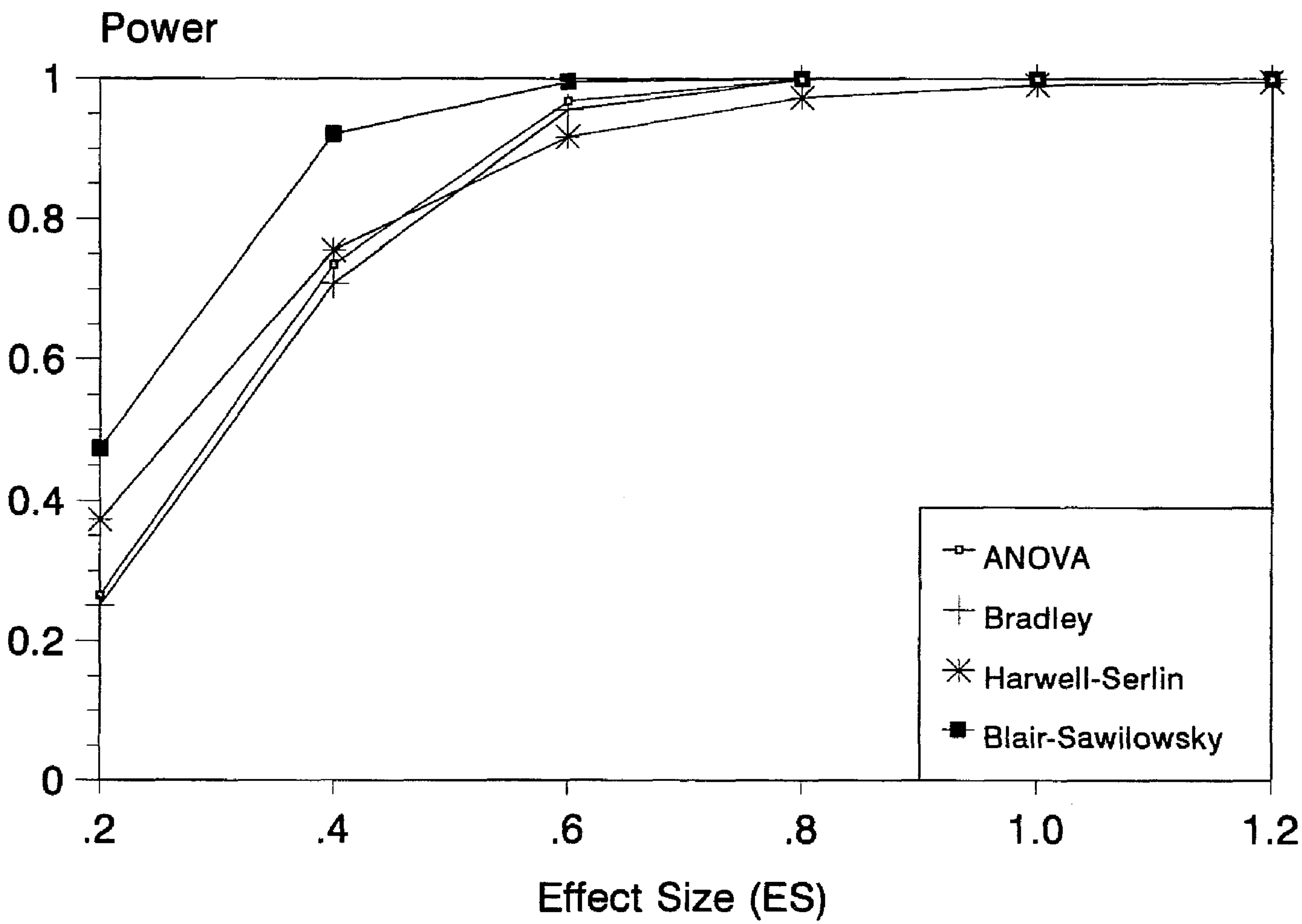

Figure 76. Comparative power of the (ab) interation for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=21.

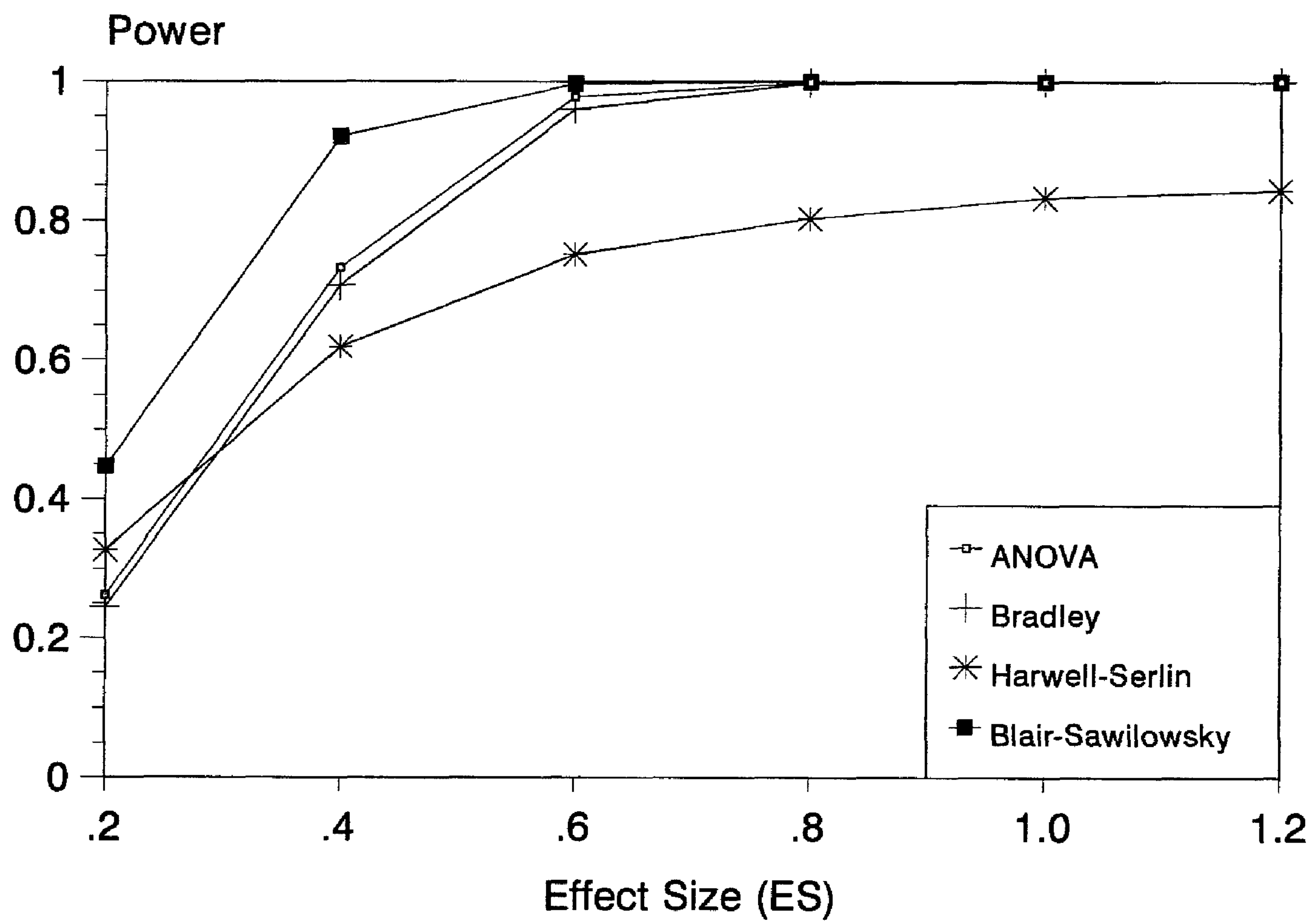

Figure 77. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=21.

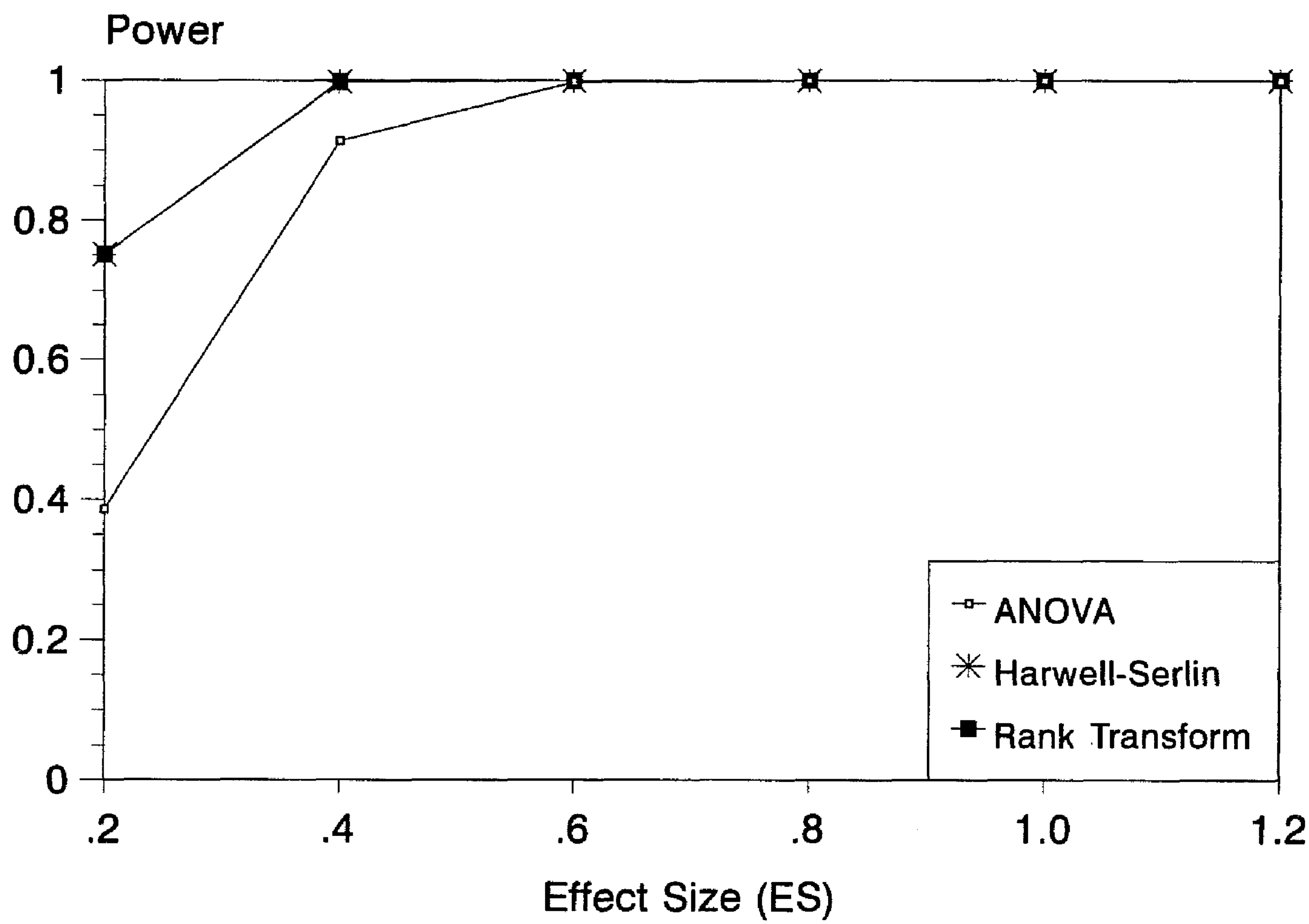

Figure 78. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the exponential distribution, alpha=.05 and n=35.

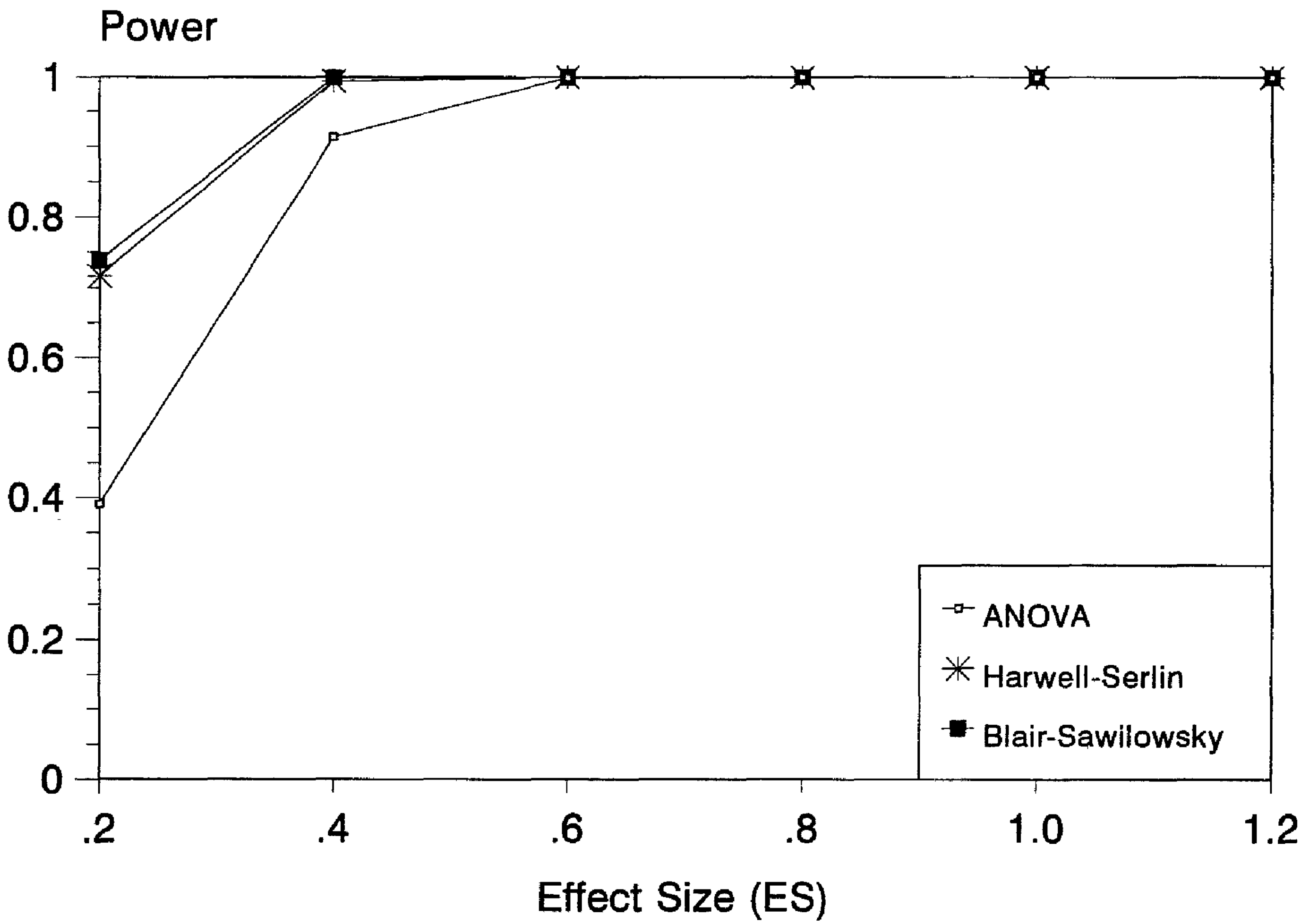

Figure 79. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=35.

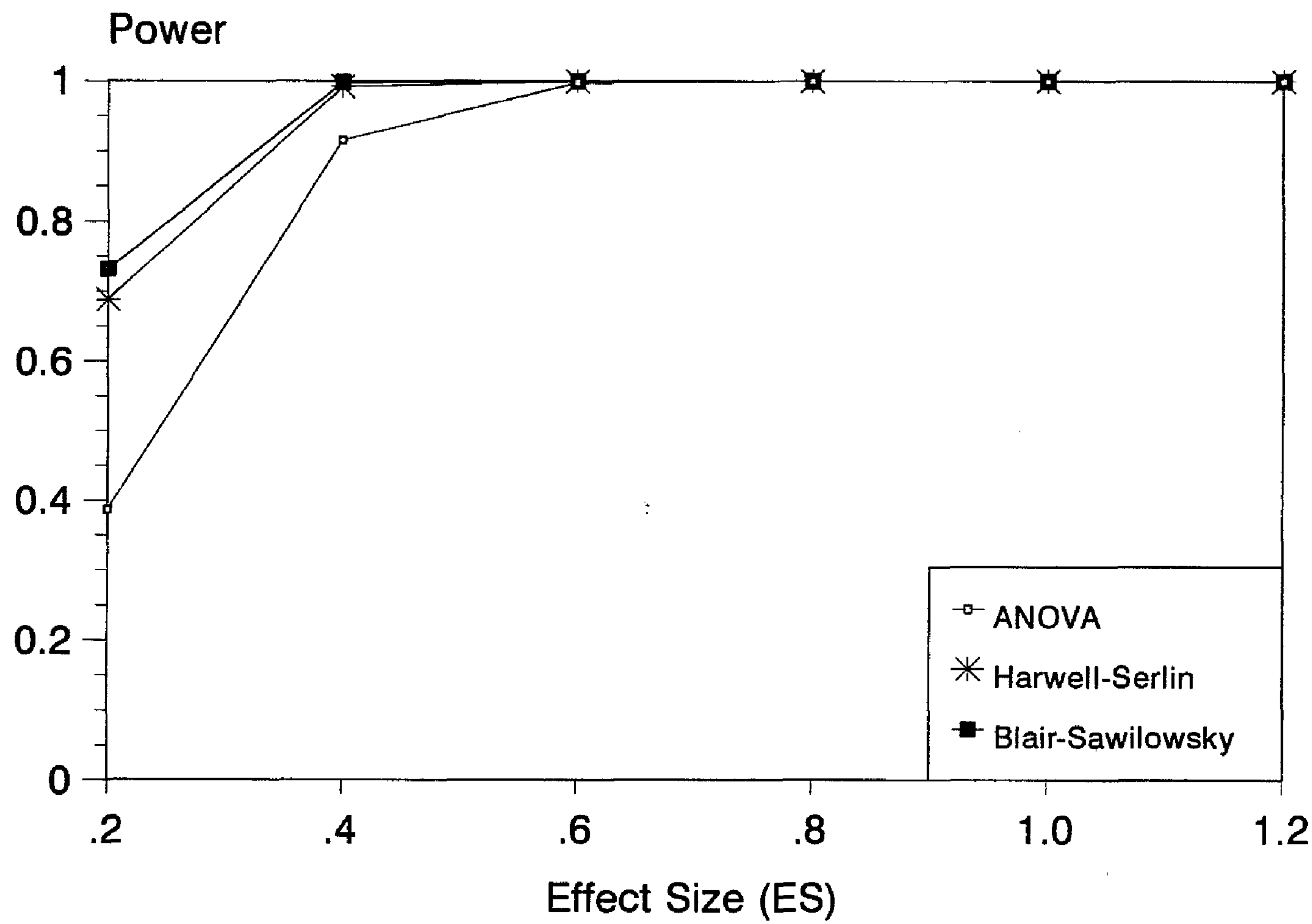

Figure 80. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=35.

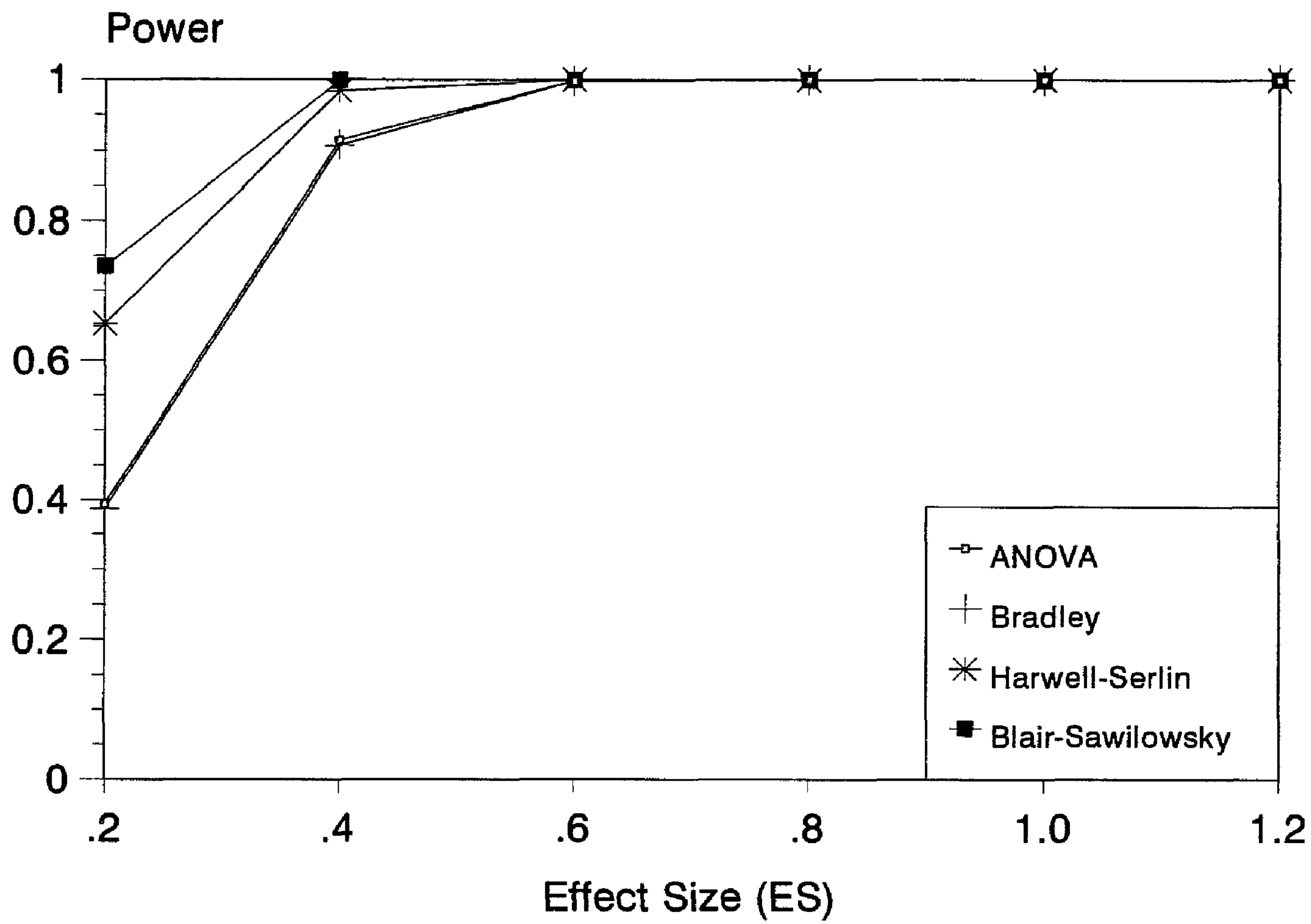

Figure 81. Comparative power of the (bc) interation for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=35.

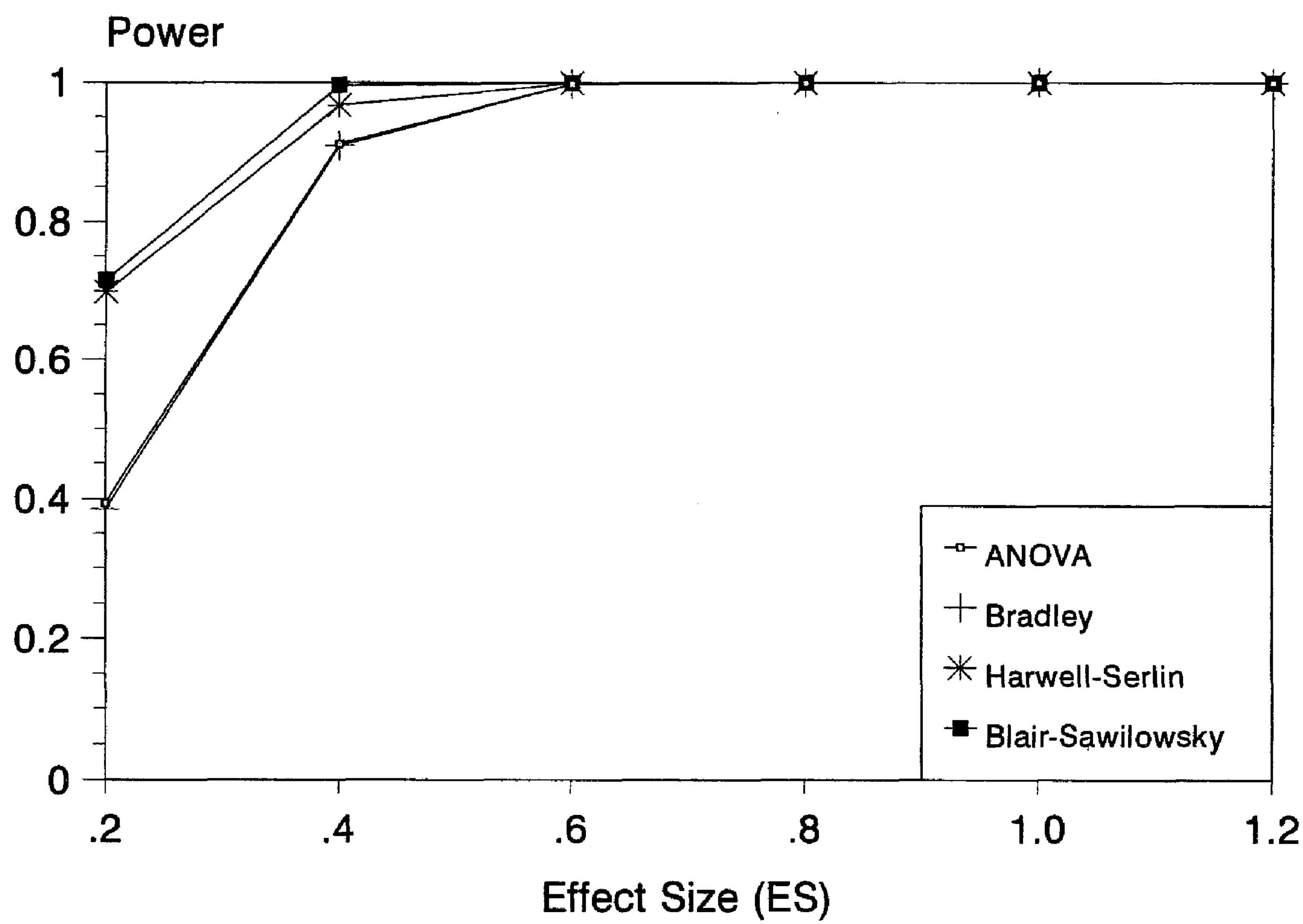

Figure 82. Comparative power of the (ac) interation for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=35.

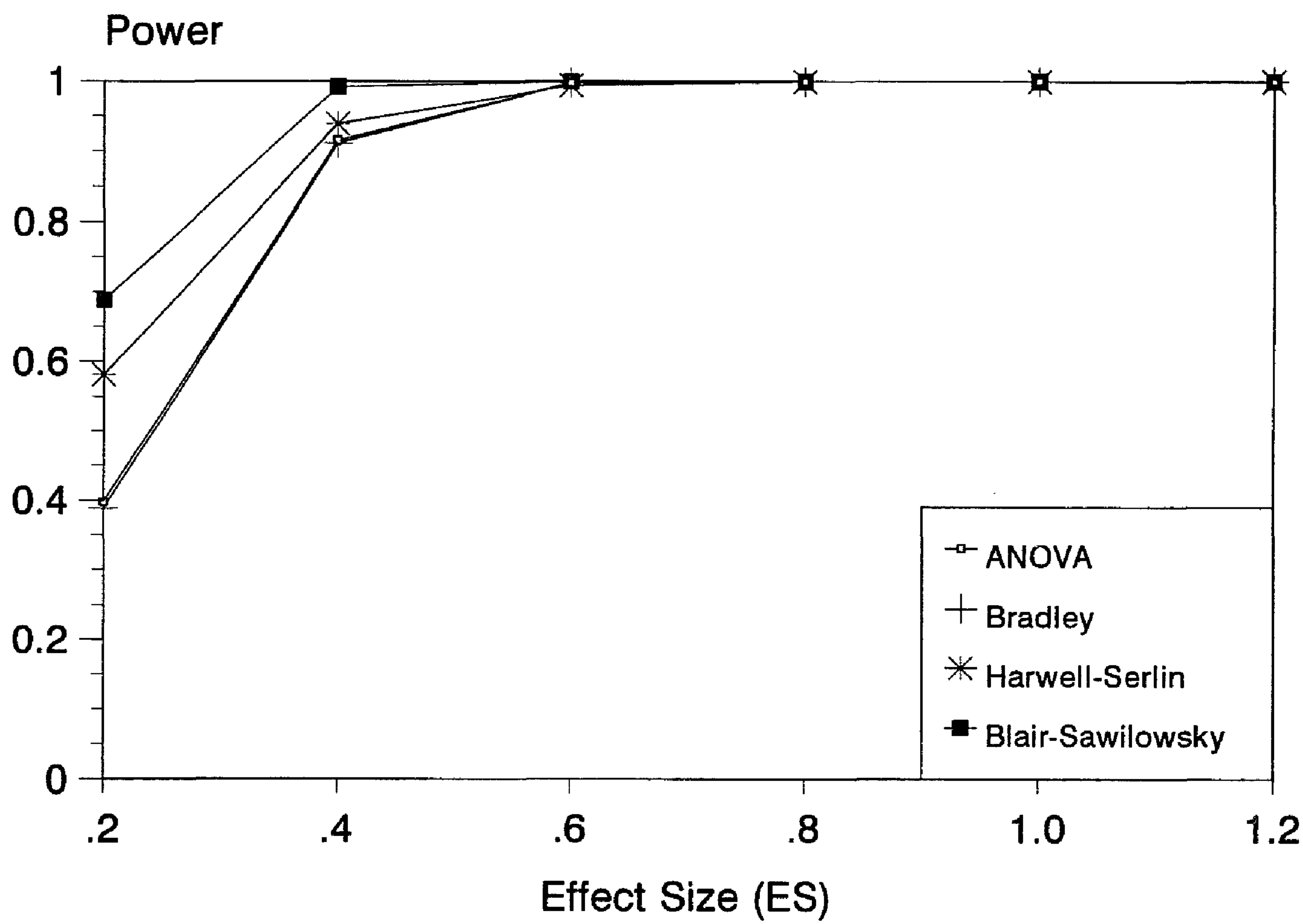

Figure 83. Comparative power of the (ab) interation for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=35.

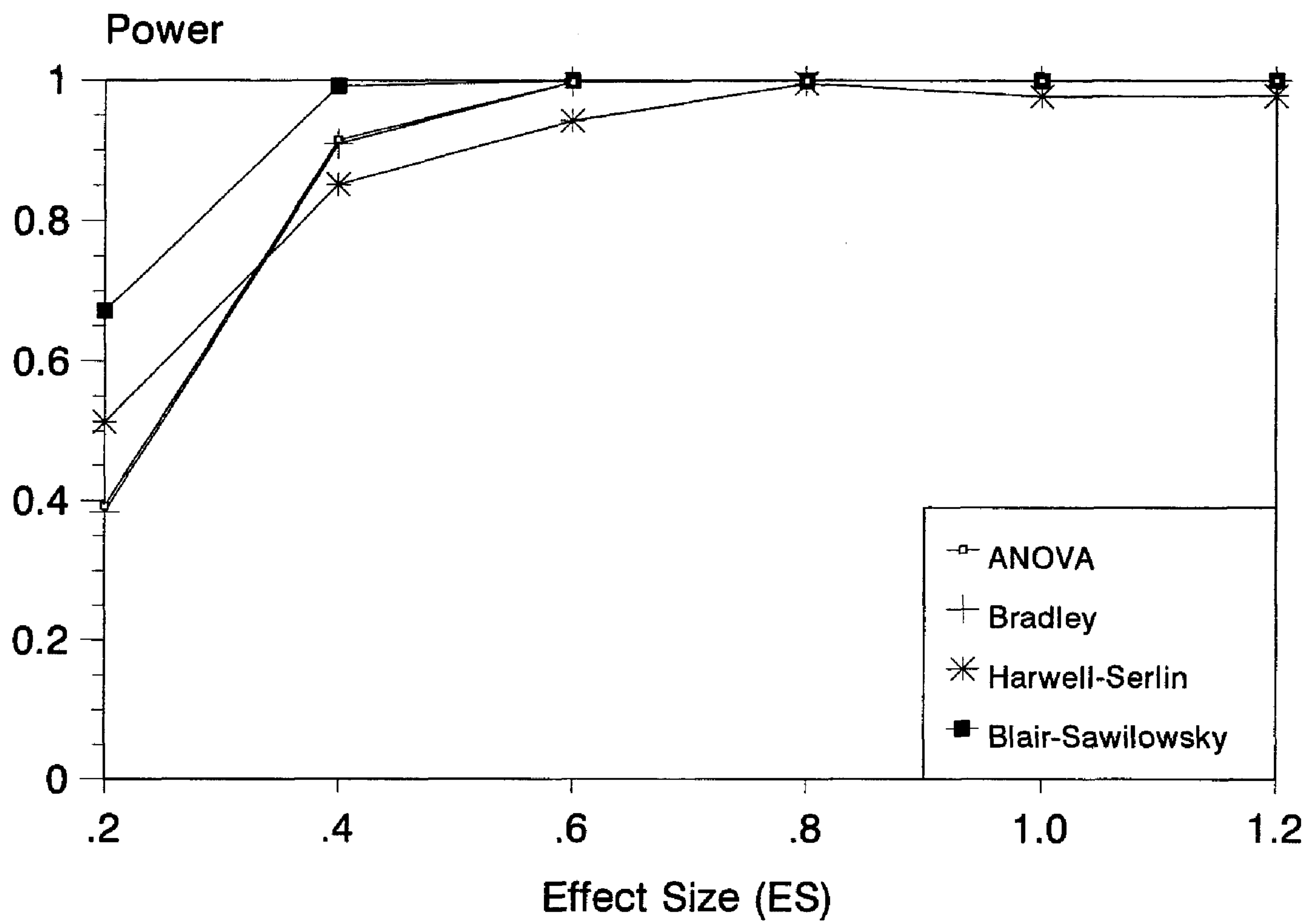

Figure 84. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the exponential distribution, alpha=.05 and n=35.

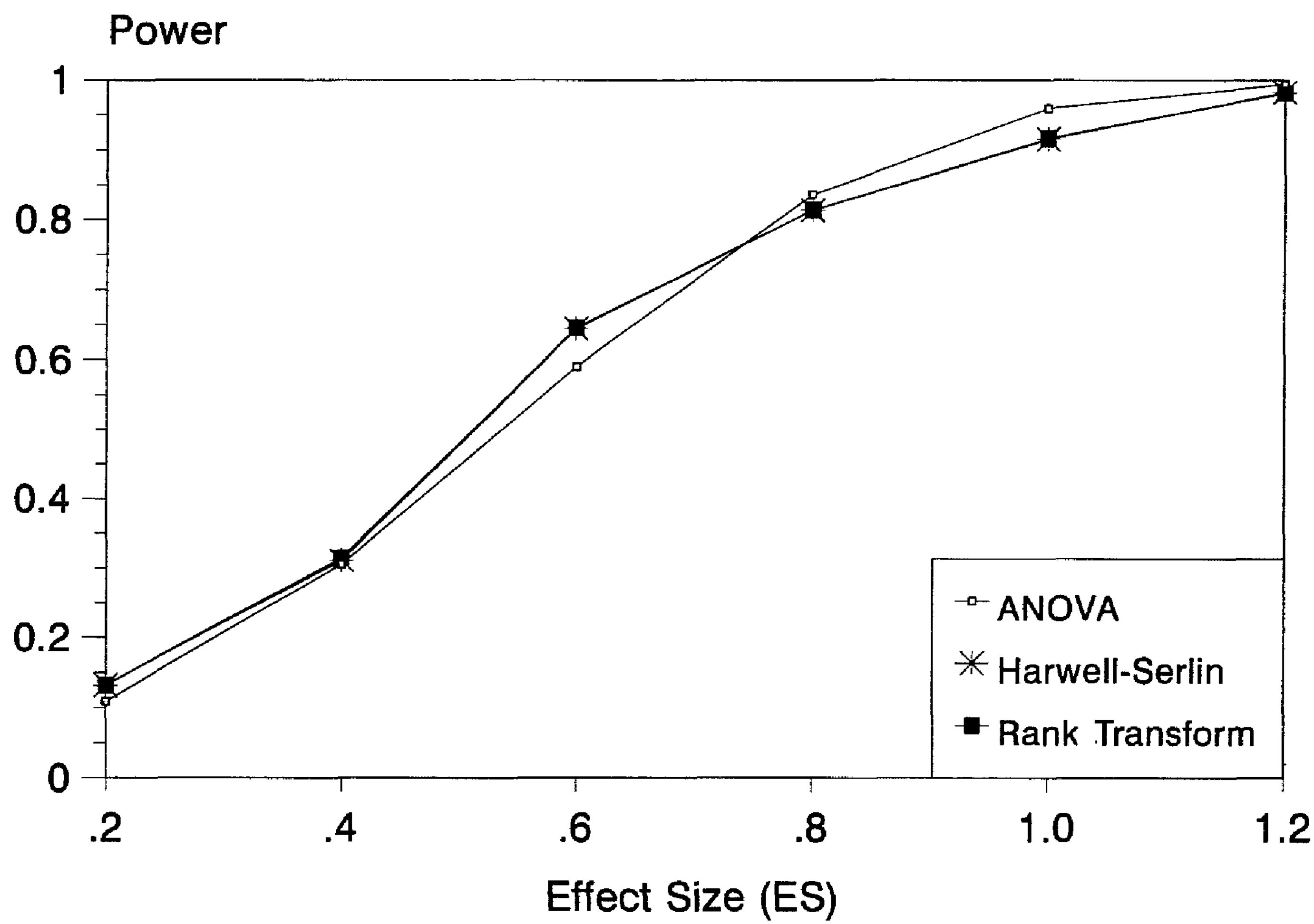

Figure 85. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=7.

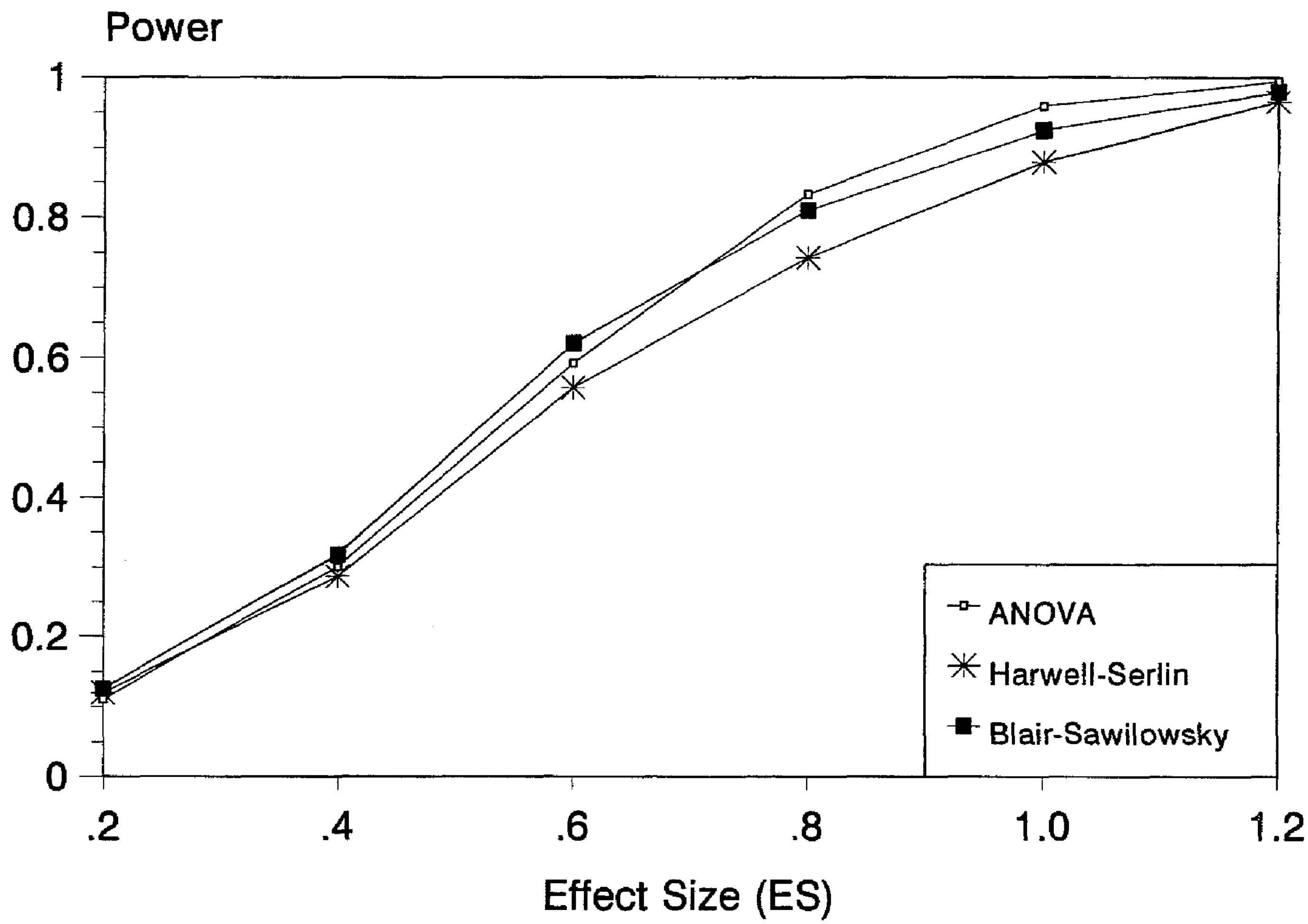

Figure 86. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05, and n=7.

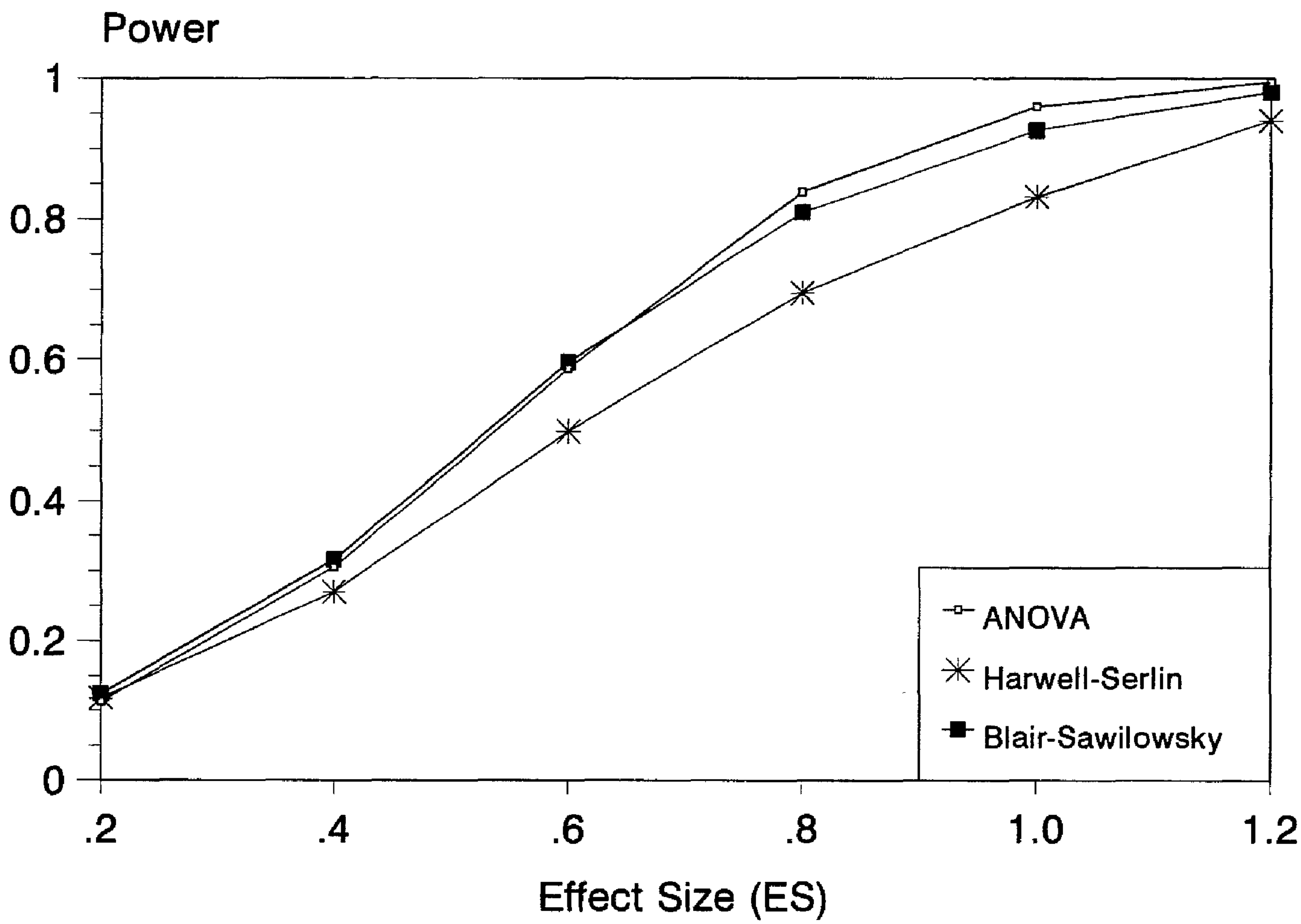

Figure 87. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=7.

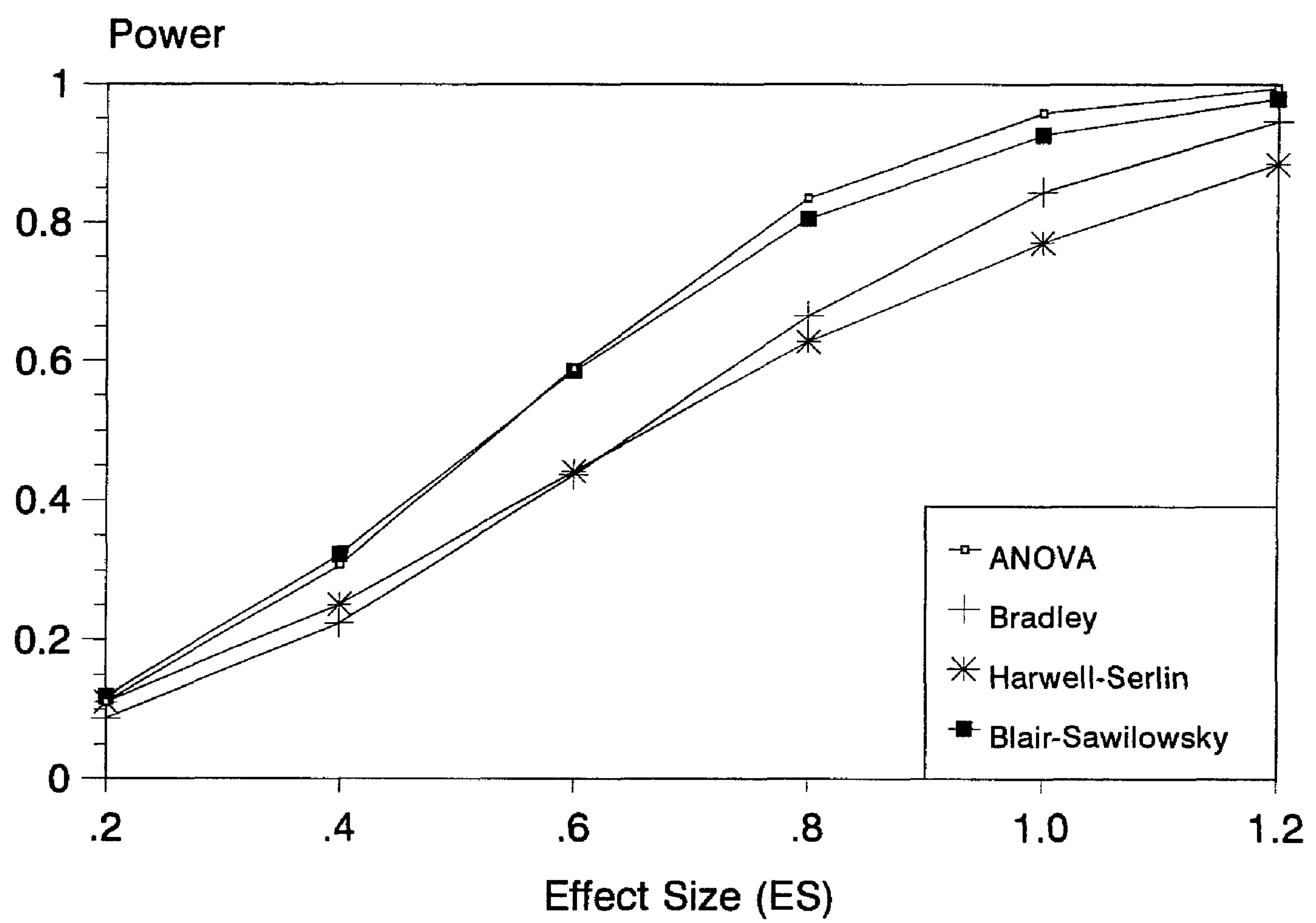

Figure 88. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=7.

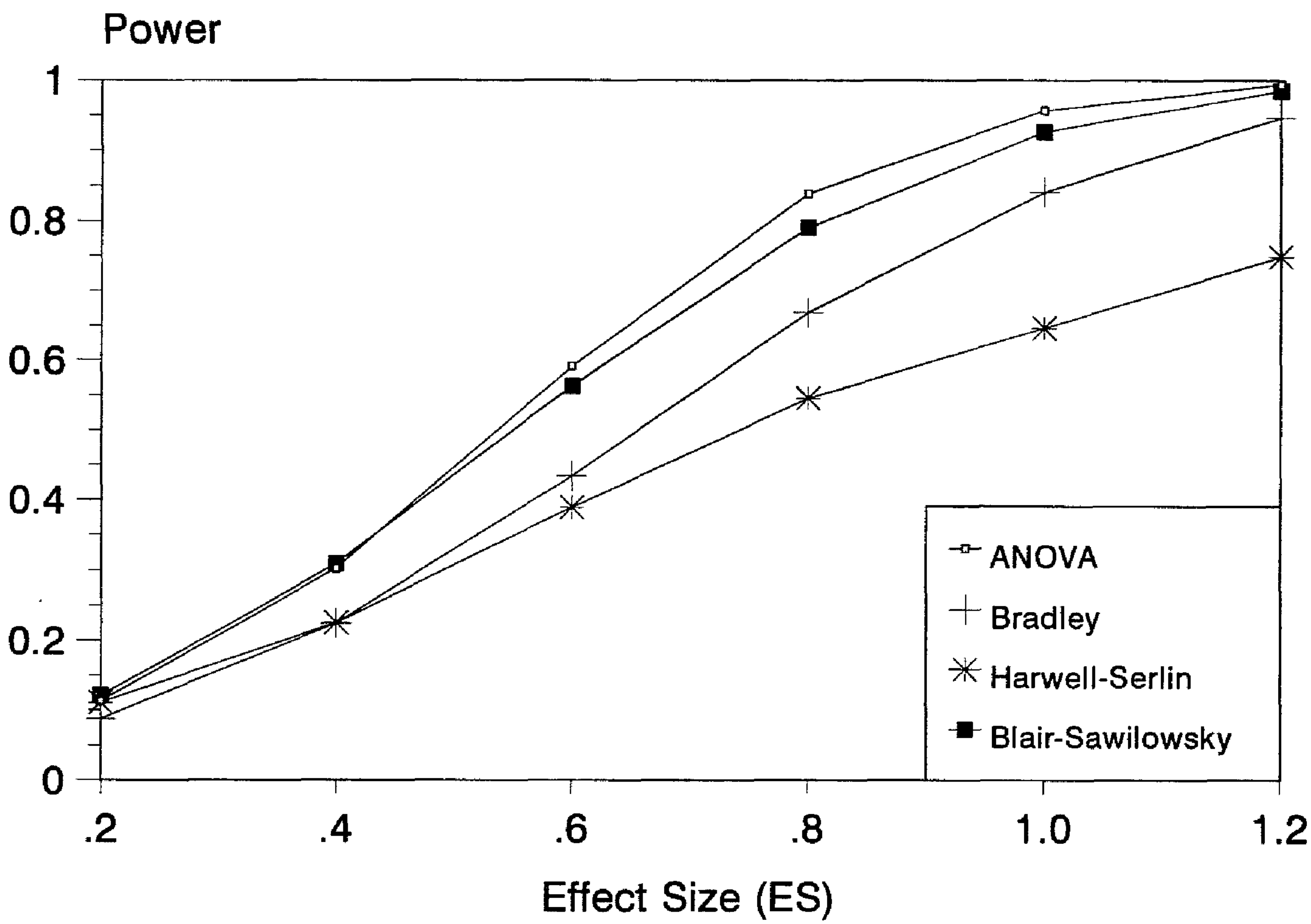

Figure 89. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=7.

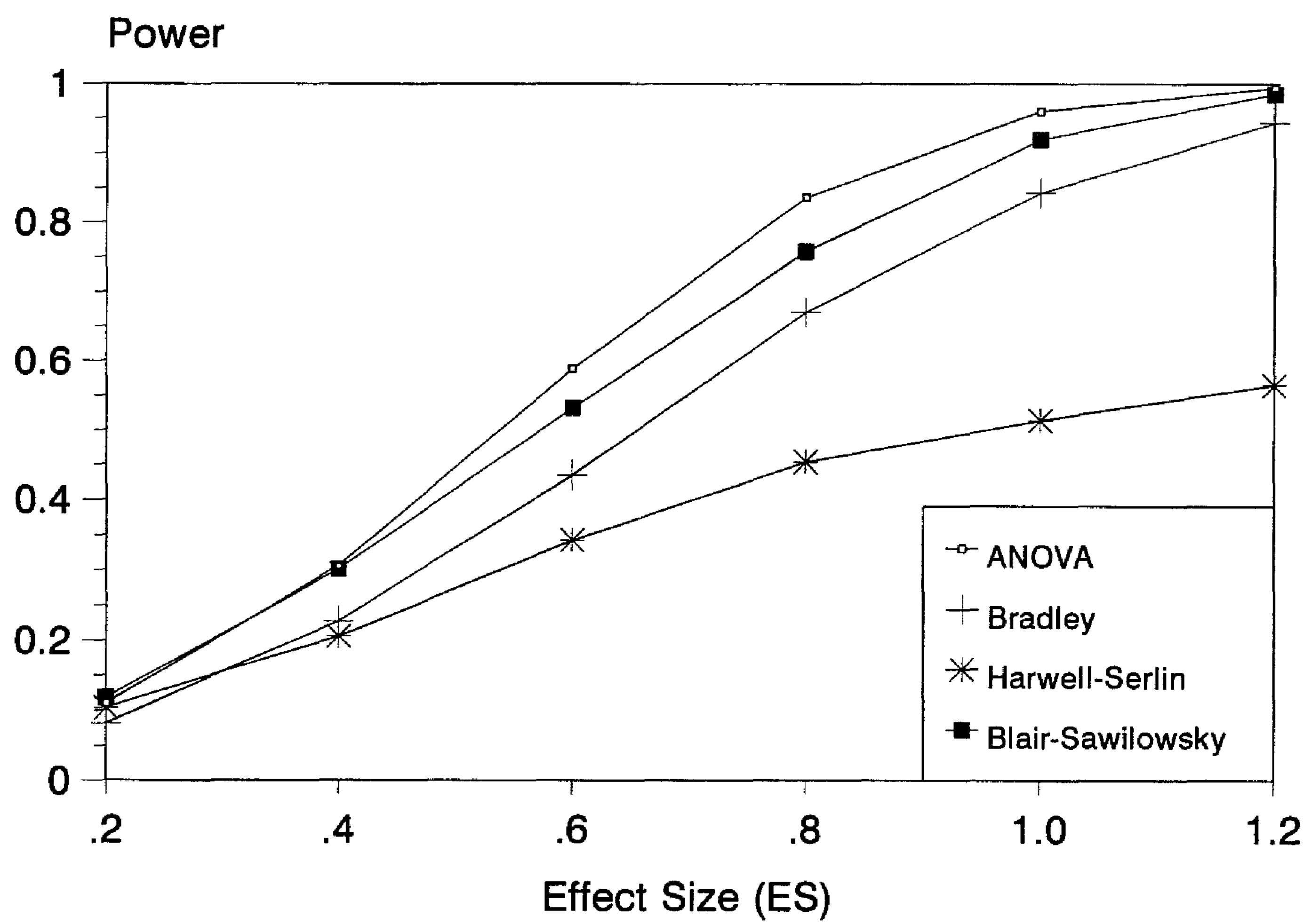

Figure 90. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=7.

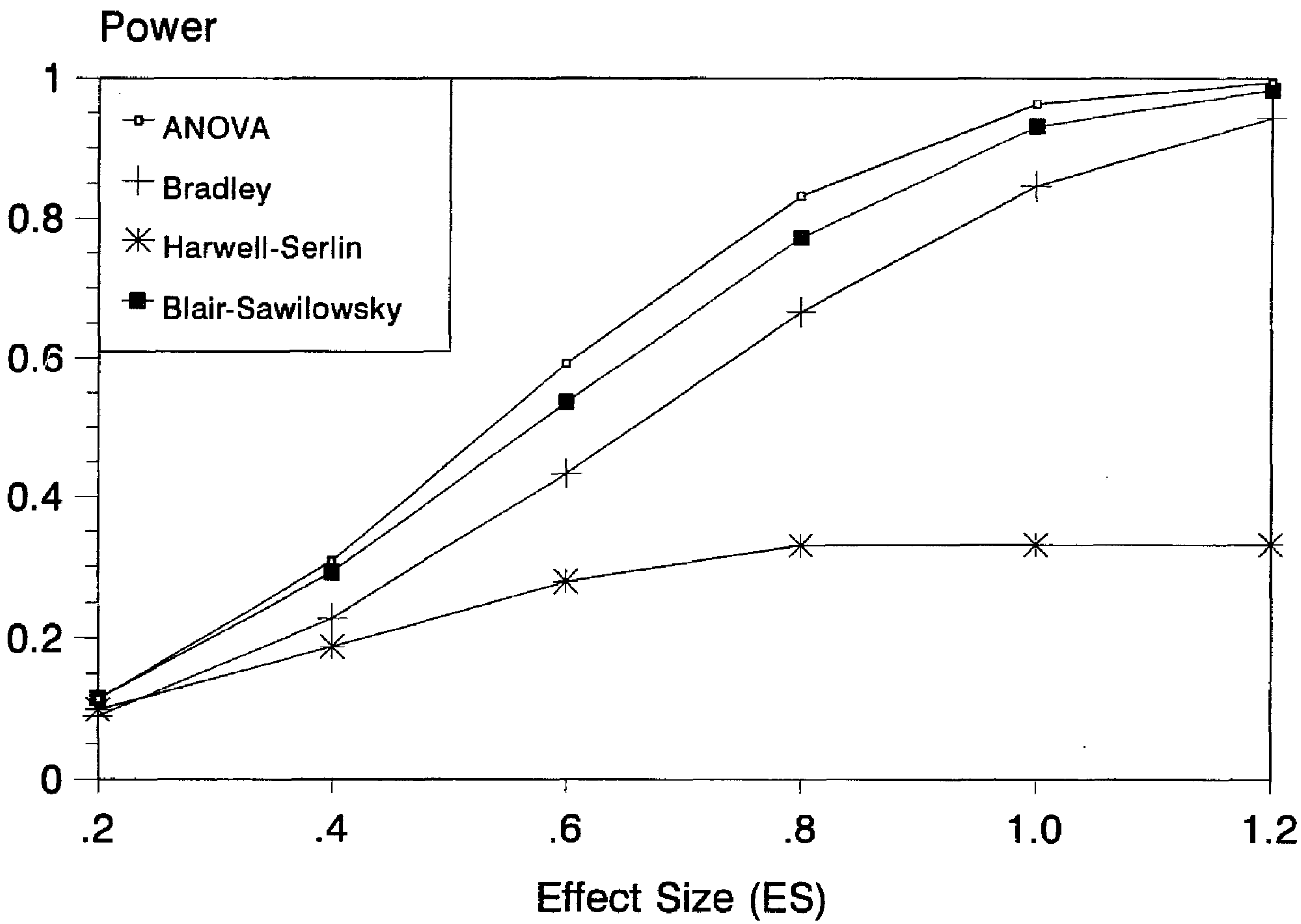

Figure 91. Comparative power of (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=7.

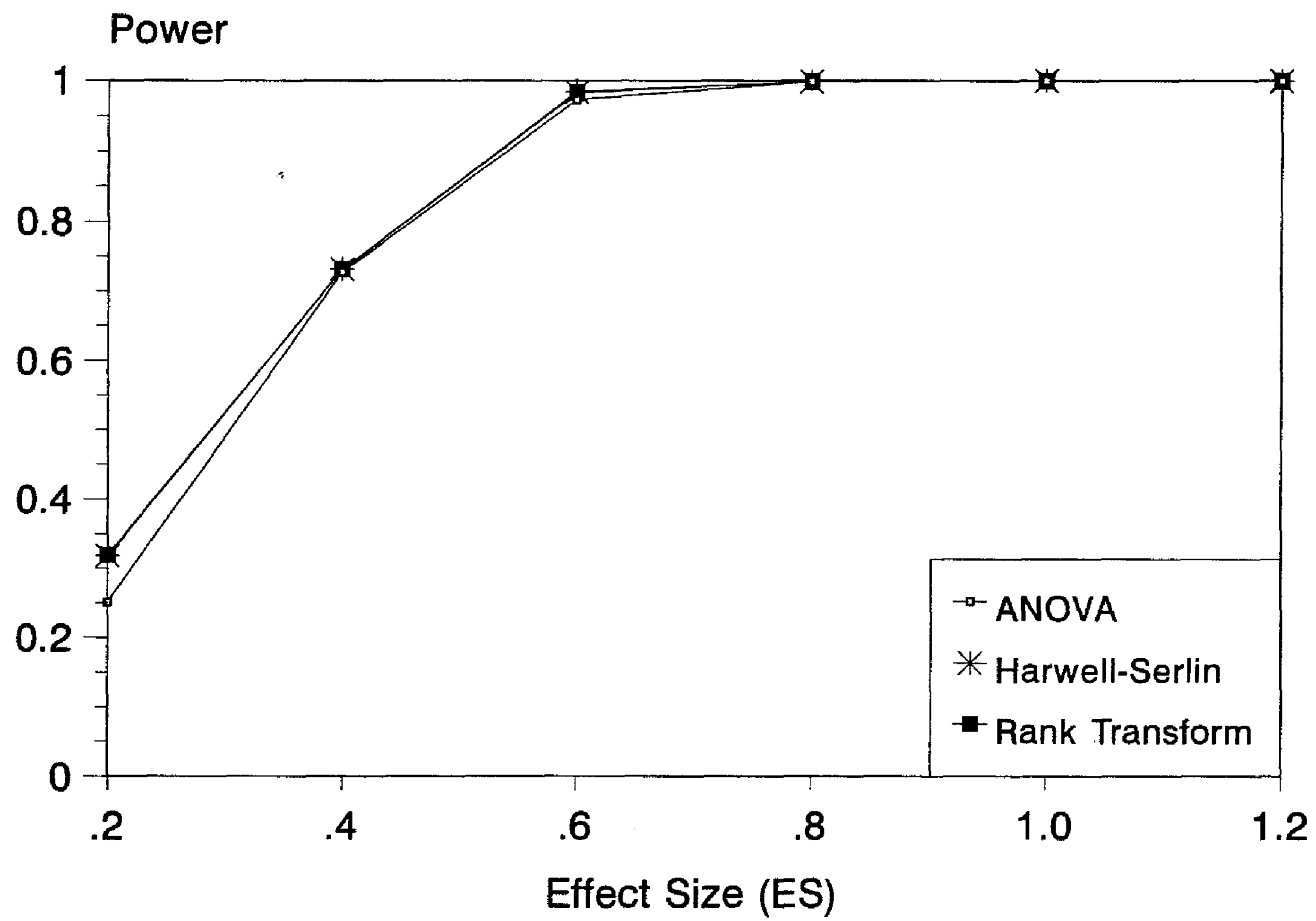

Figure 92. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=21.

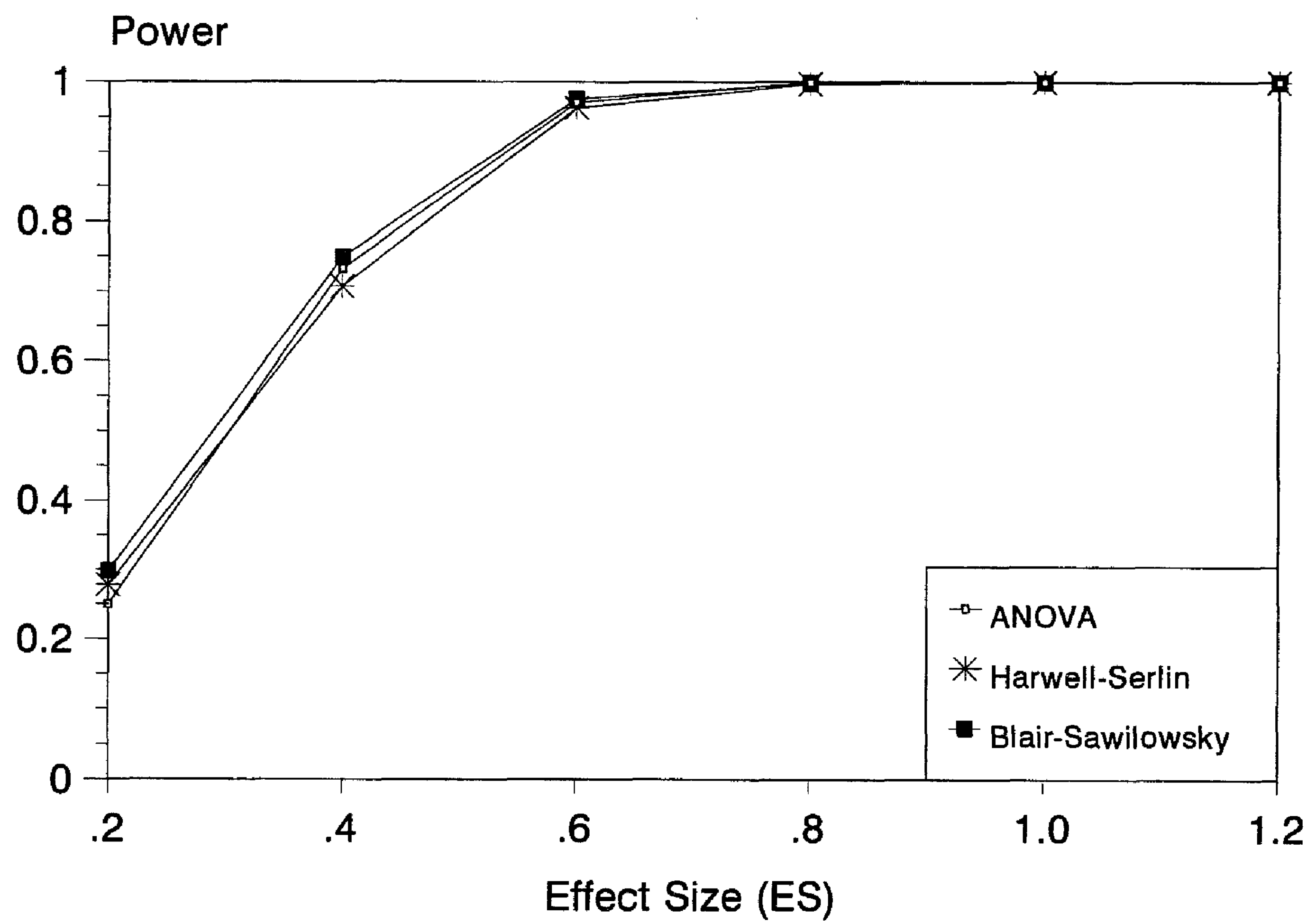

Figure 93. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=21.

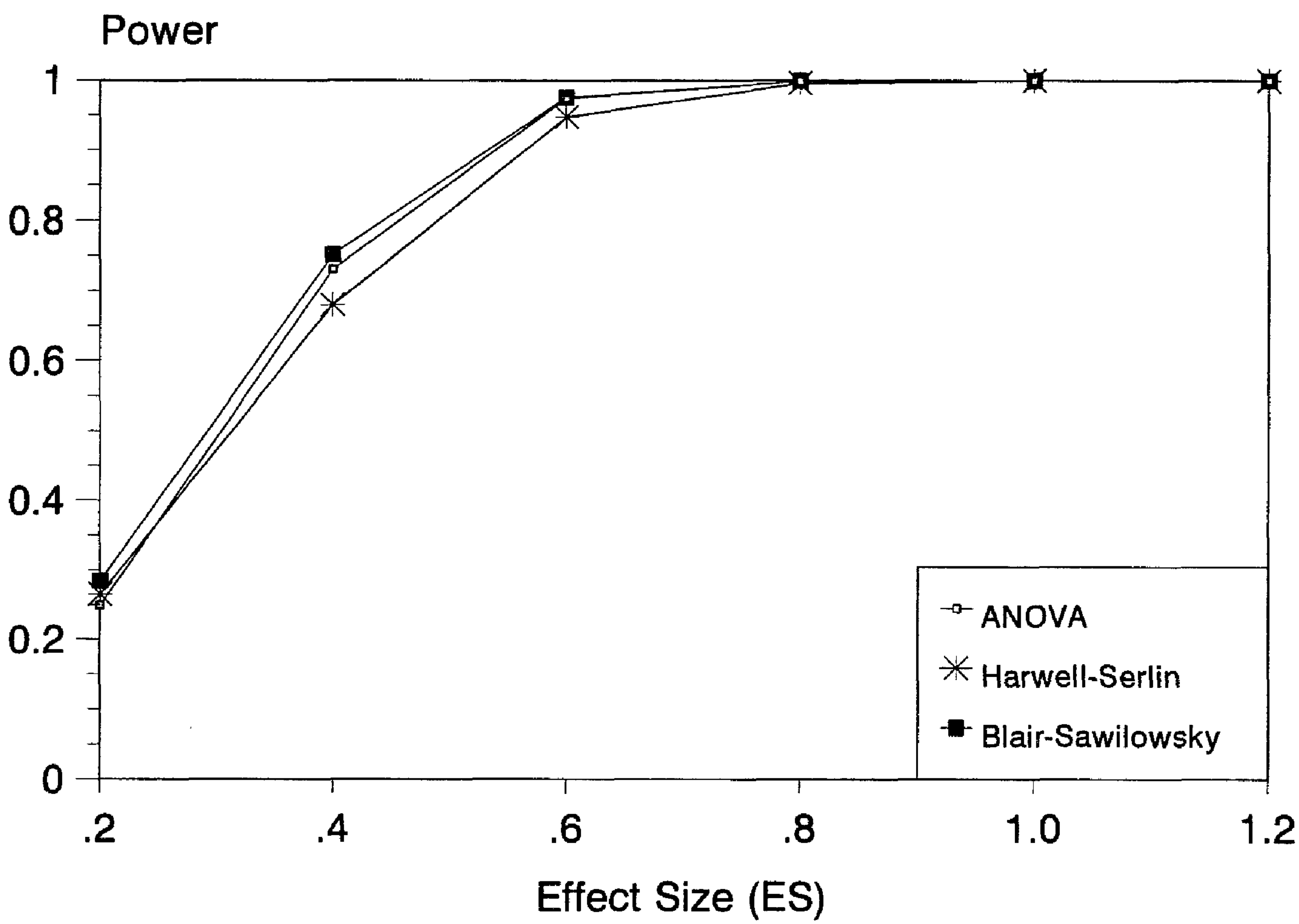

Figure 94. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=21.

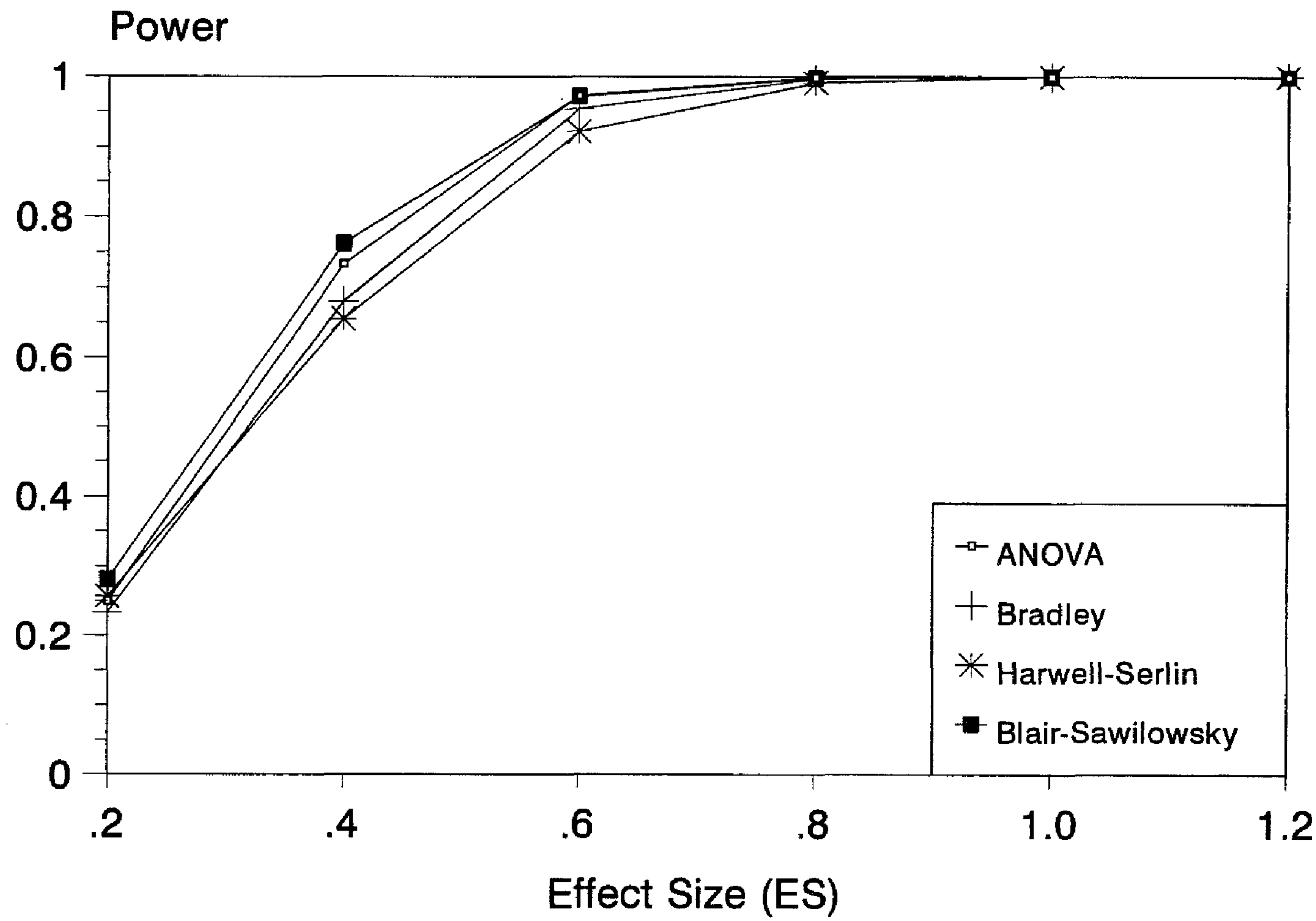

Figure 95. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=21.

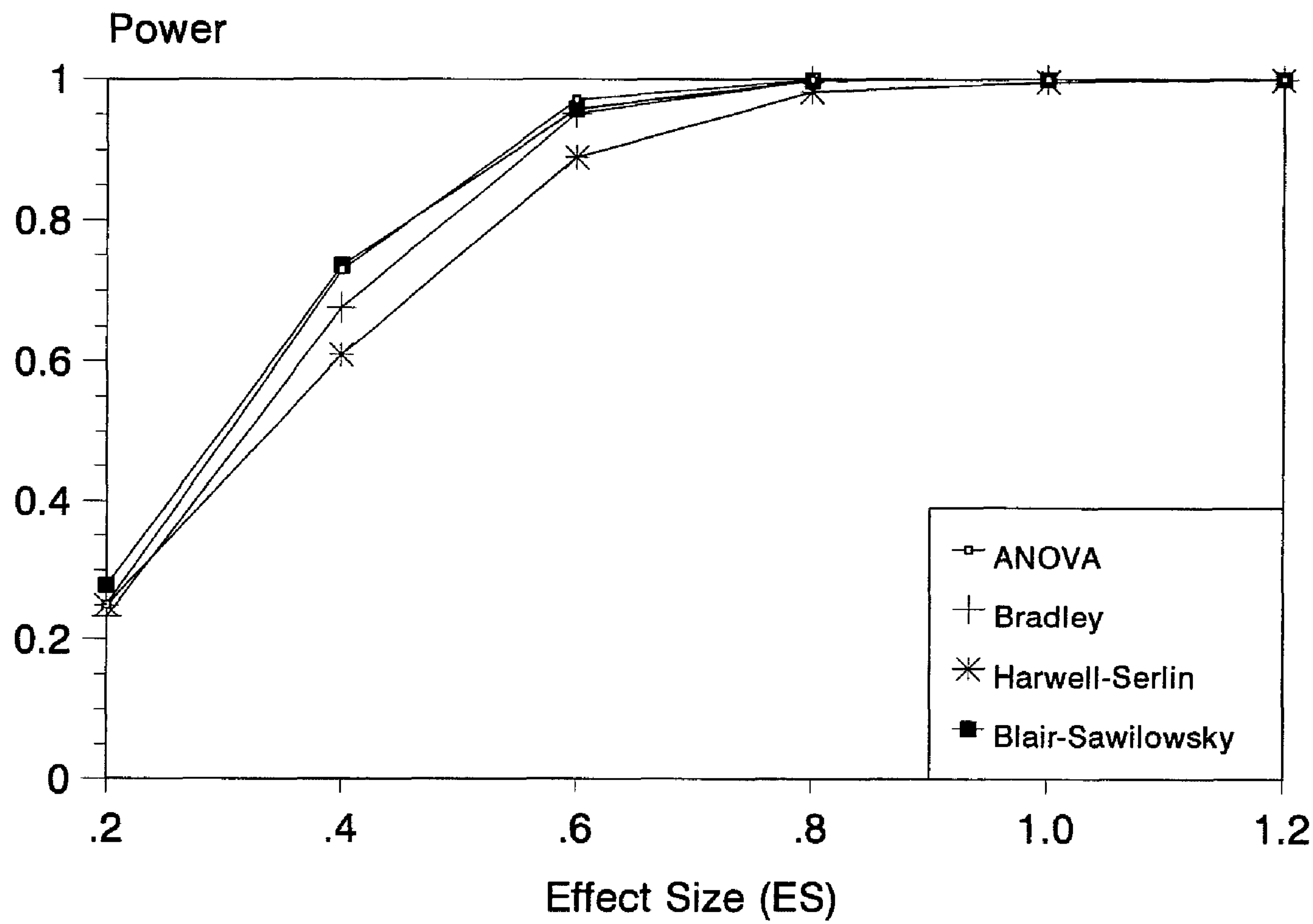

Figure 96. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=21.

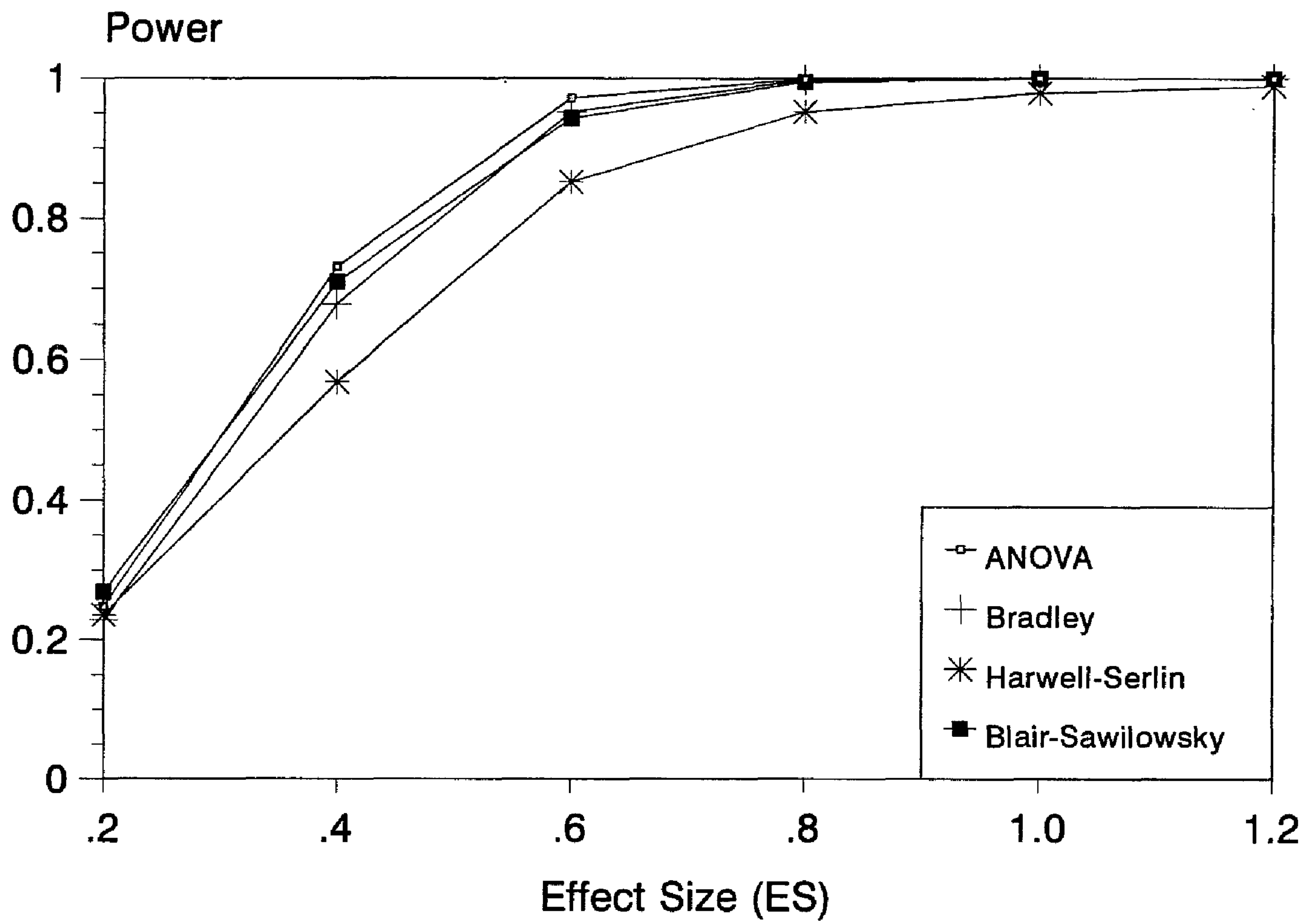

Figure 97. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=21.

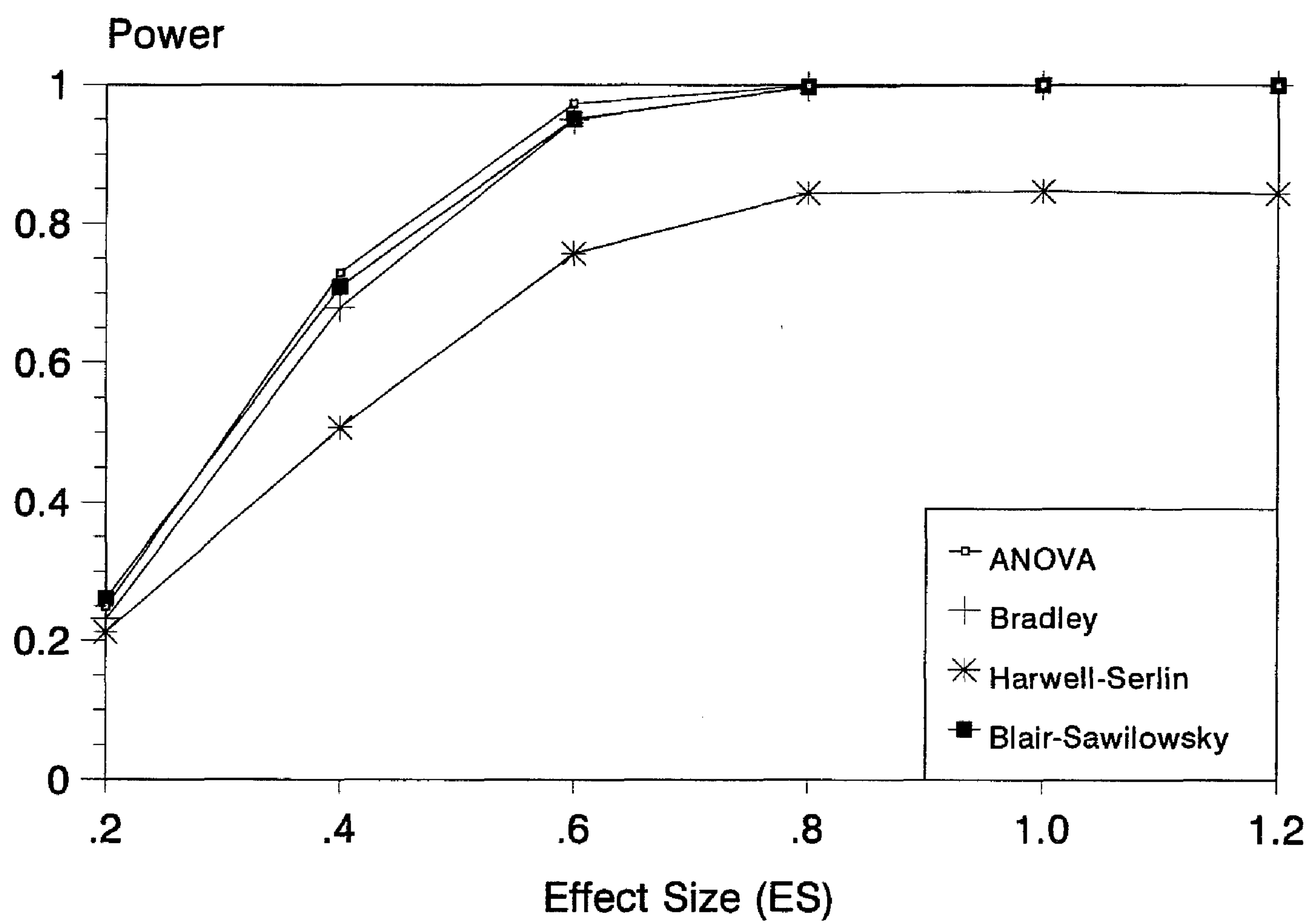

Figure 98. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=21.

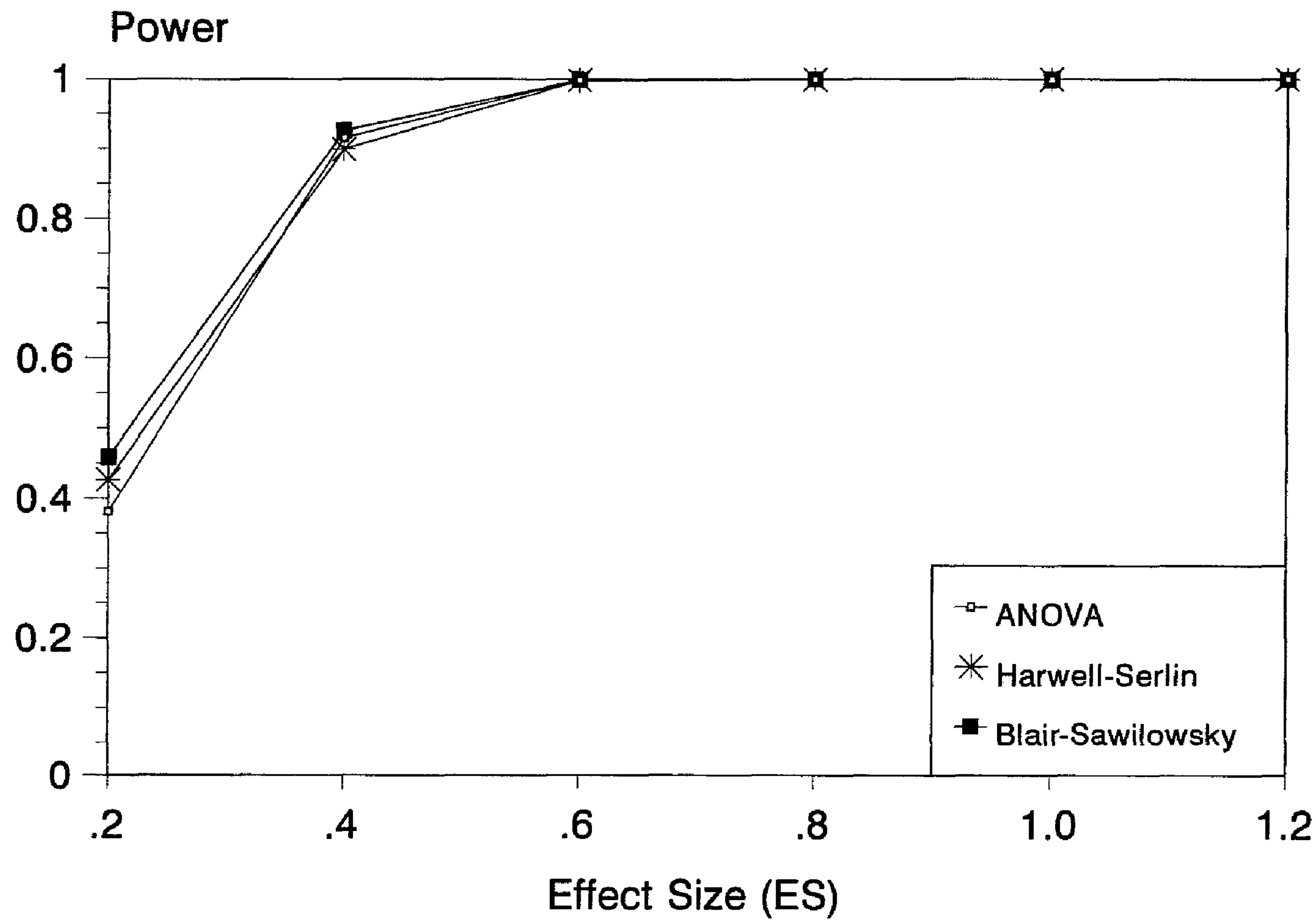

Figure 100. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=35.

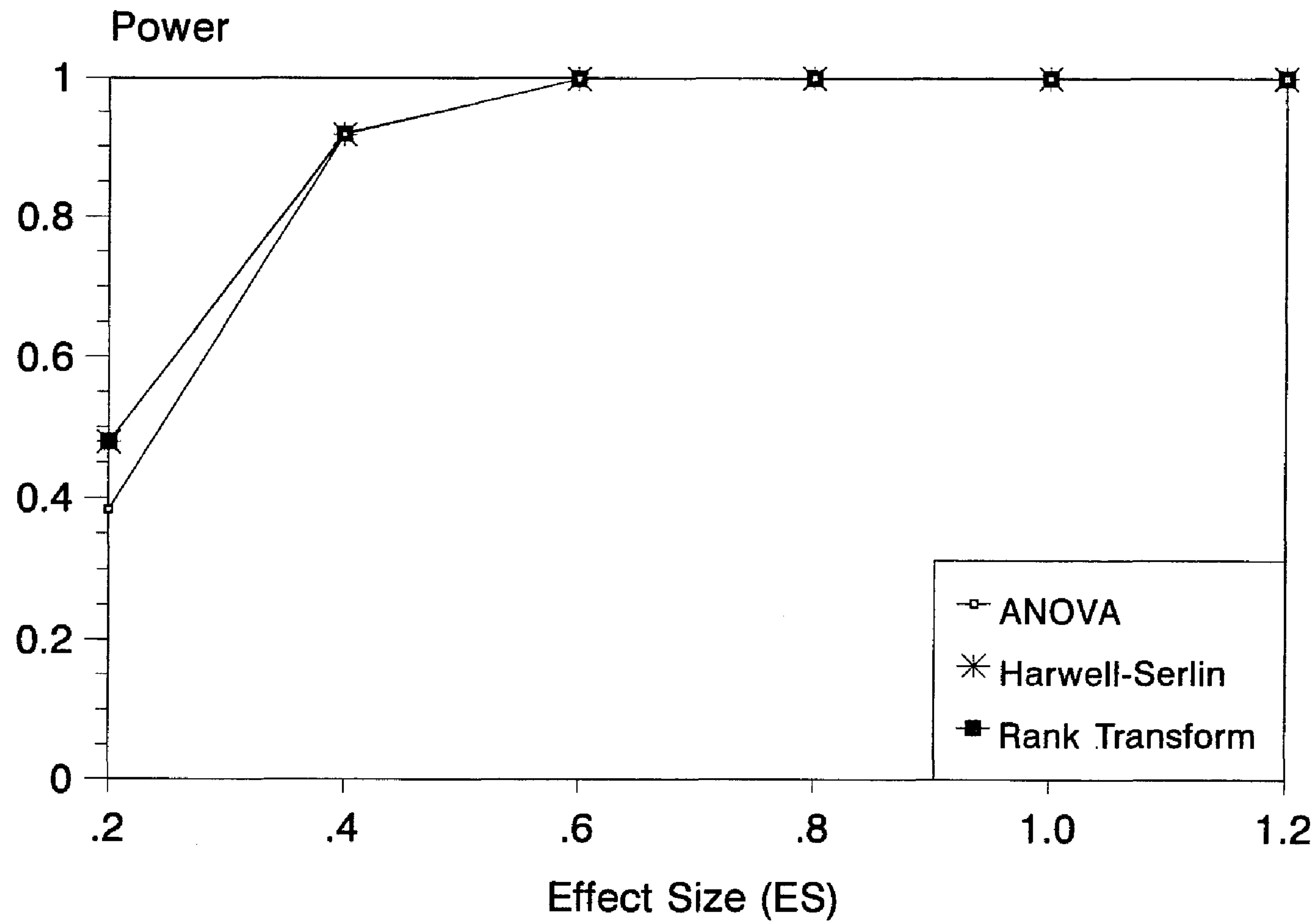

Figure 99. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=35.

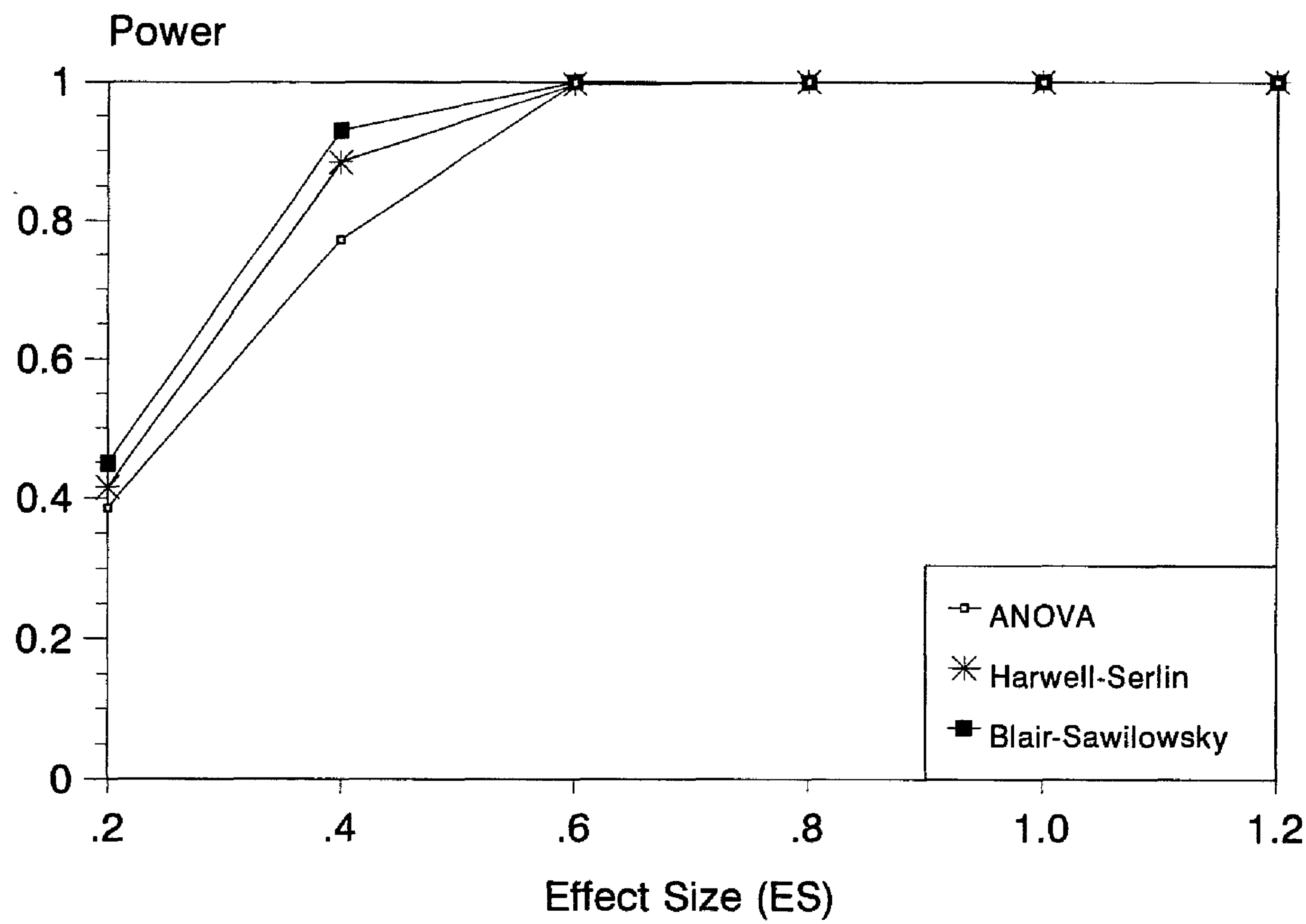

Figure 101. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=35.

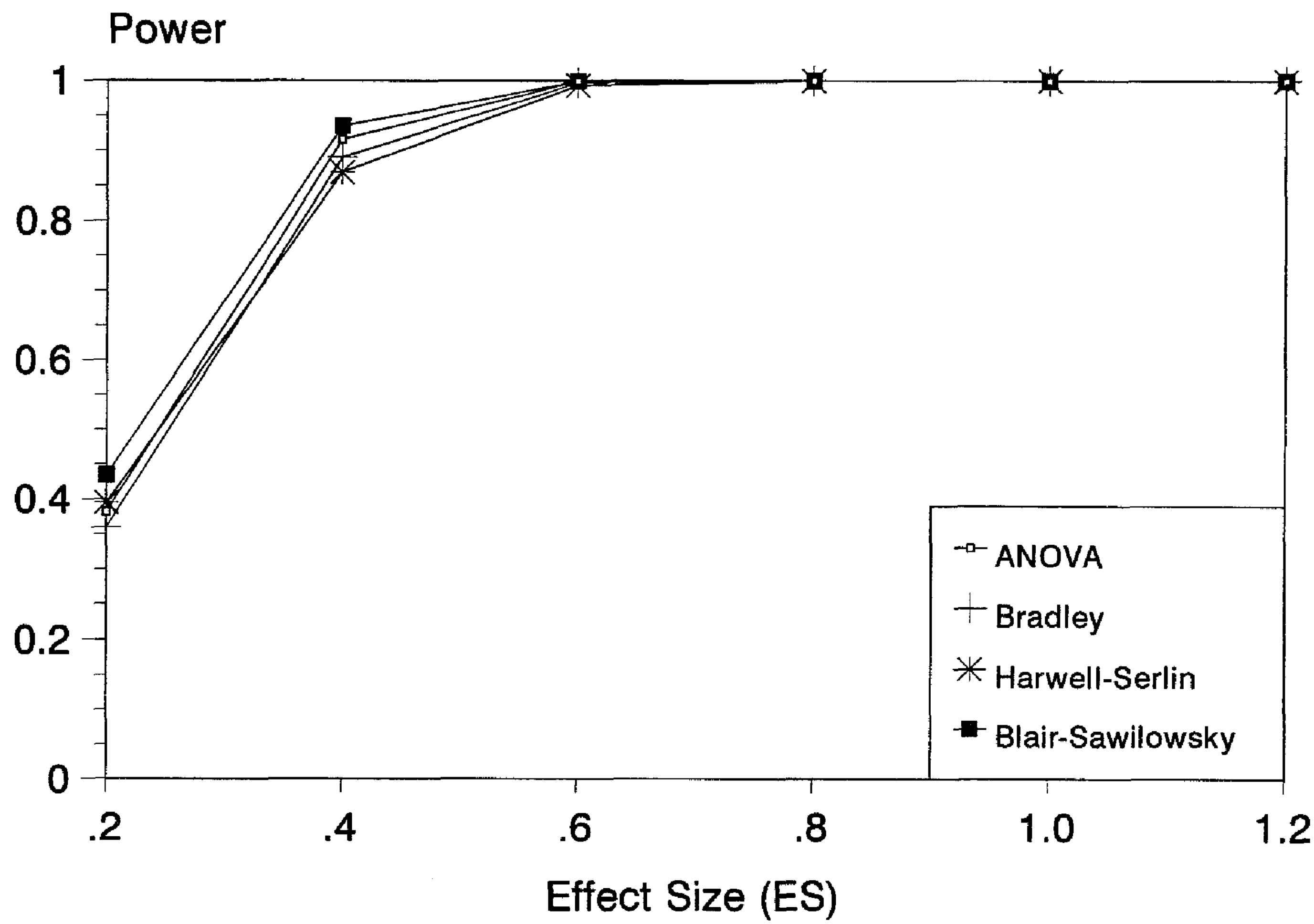

Figure 102. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=35.

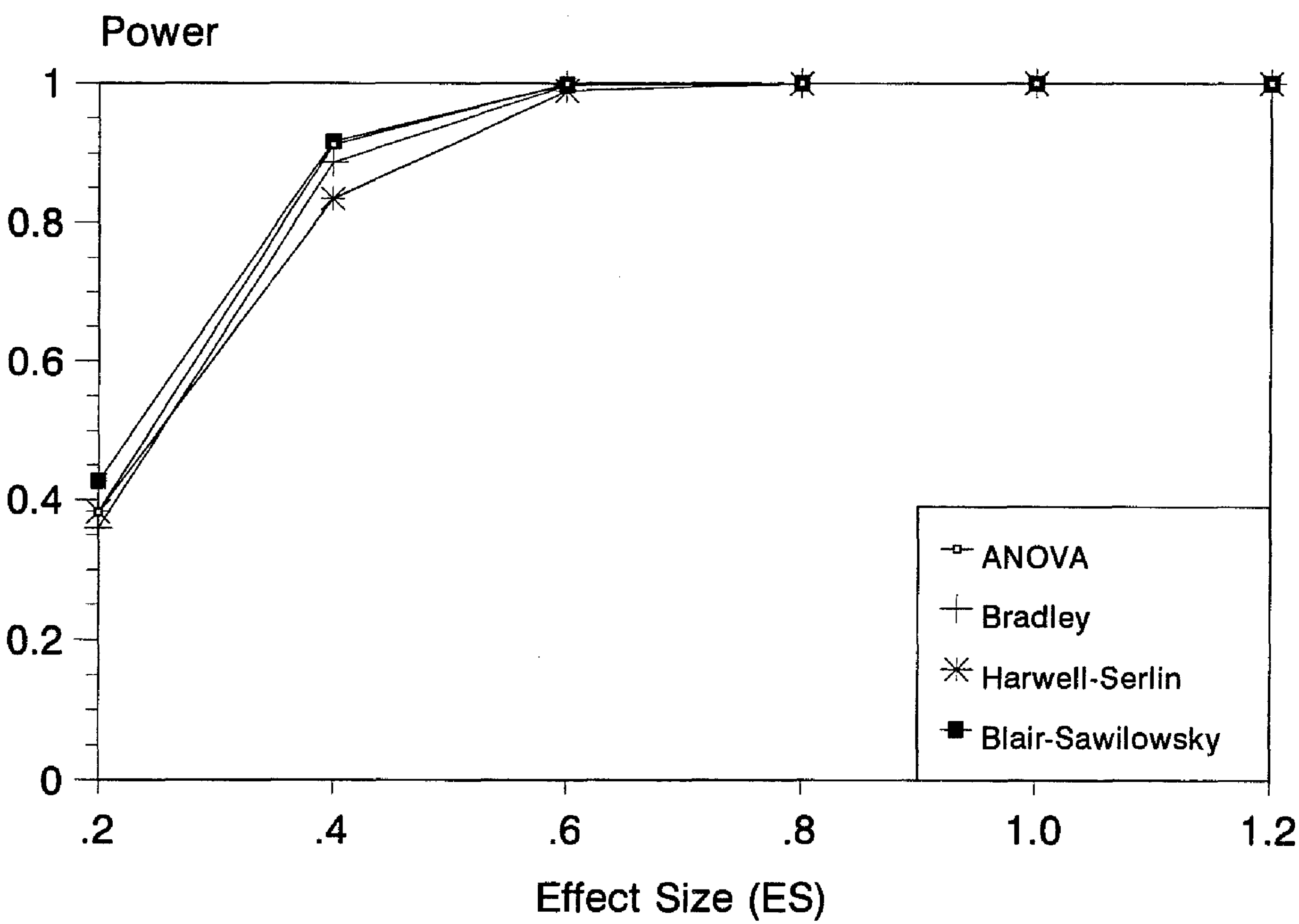

Figure 103. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=35.

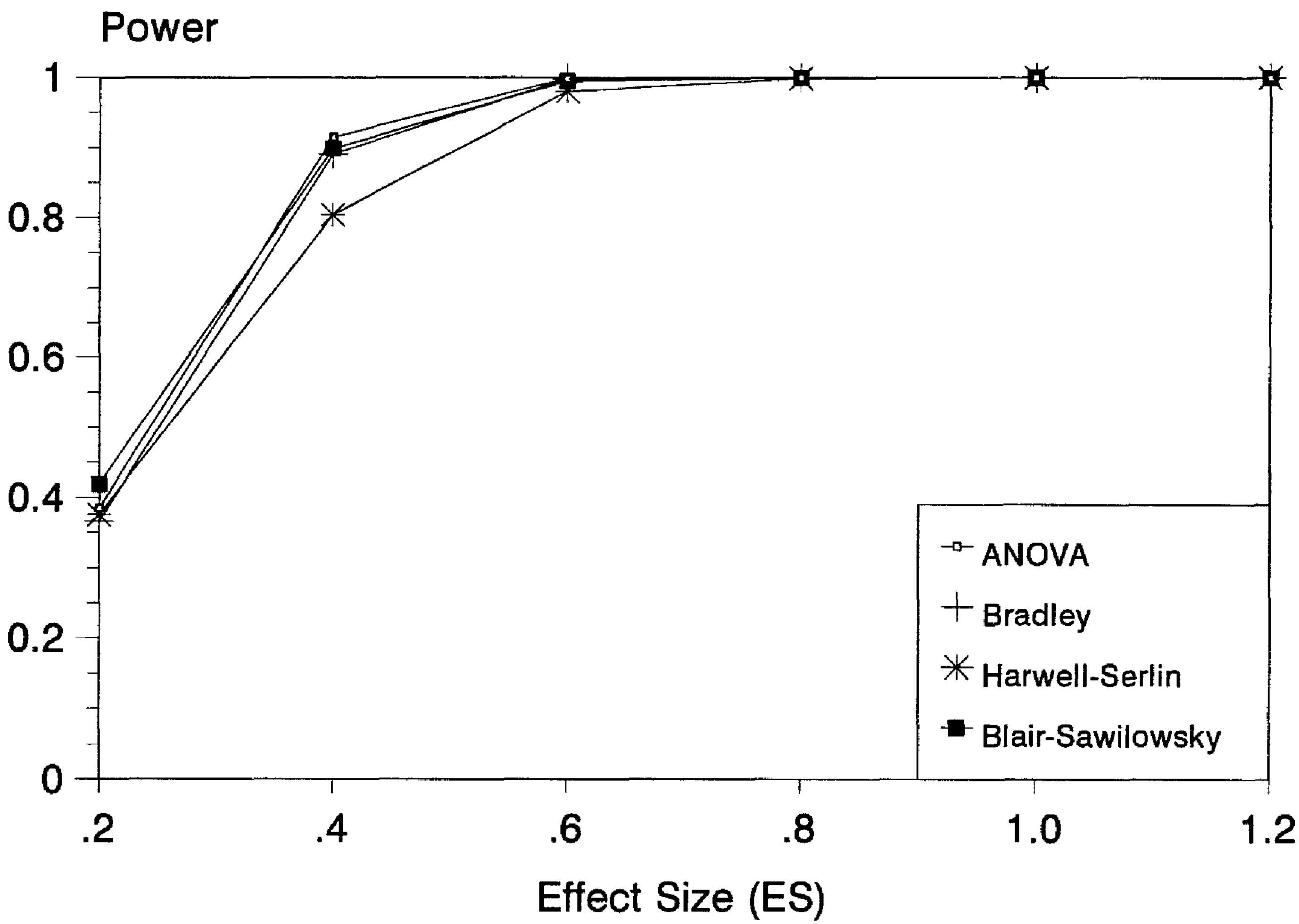

Figure 104. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=35.

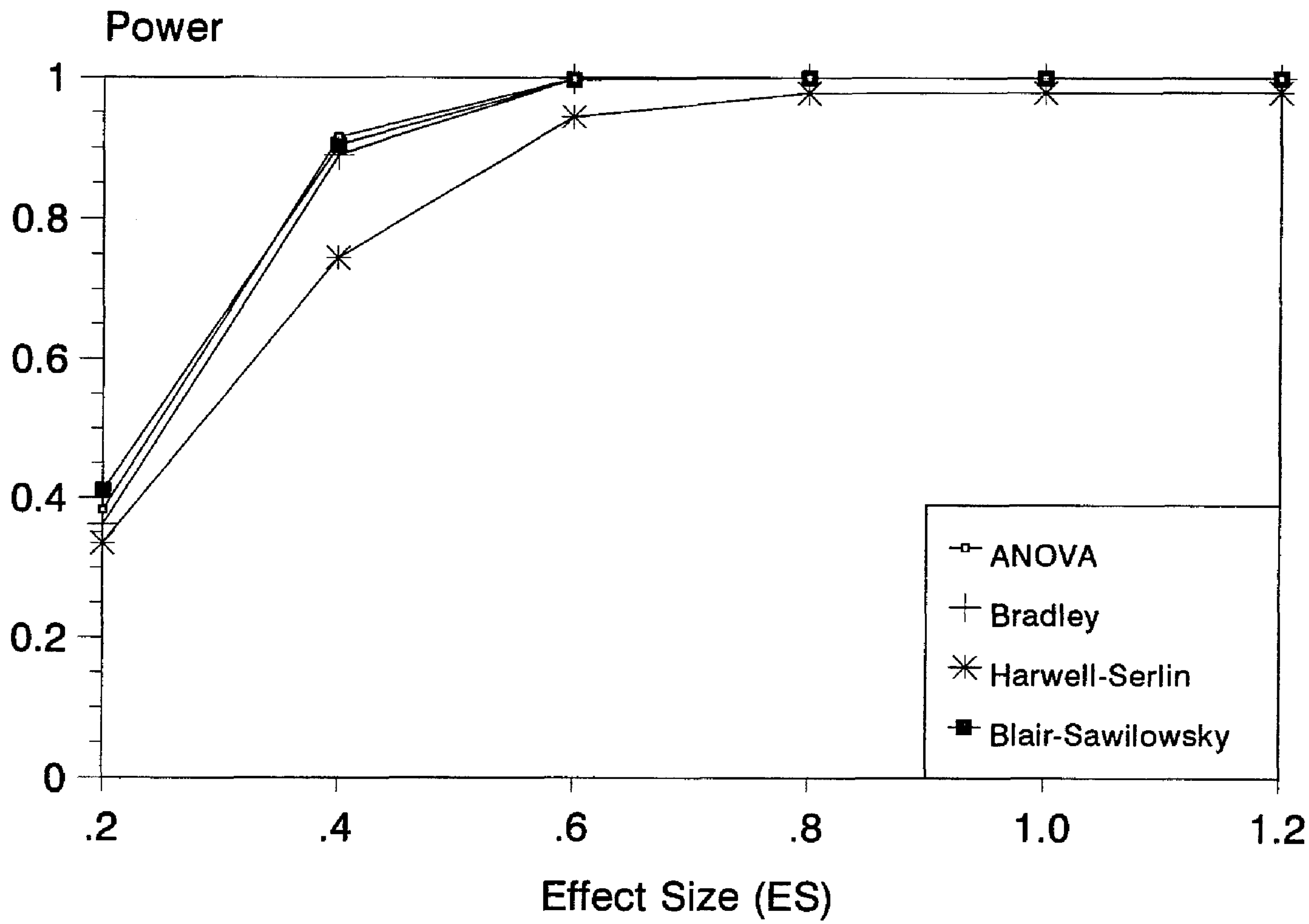

Figure 105. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the multi-modal lumpy data set, alpha=.05 and n=35.

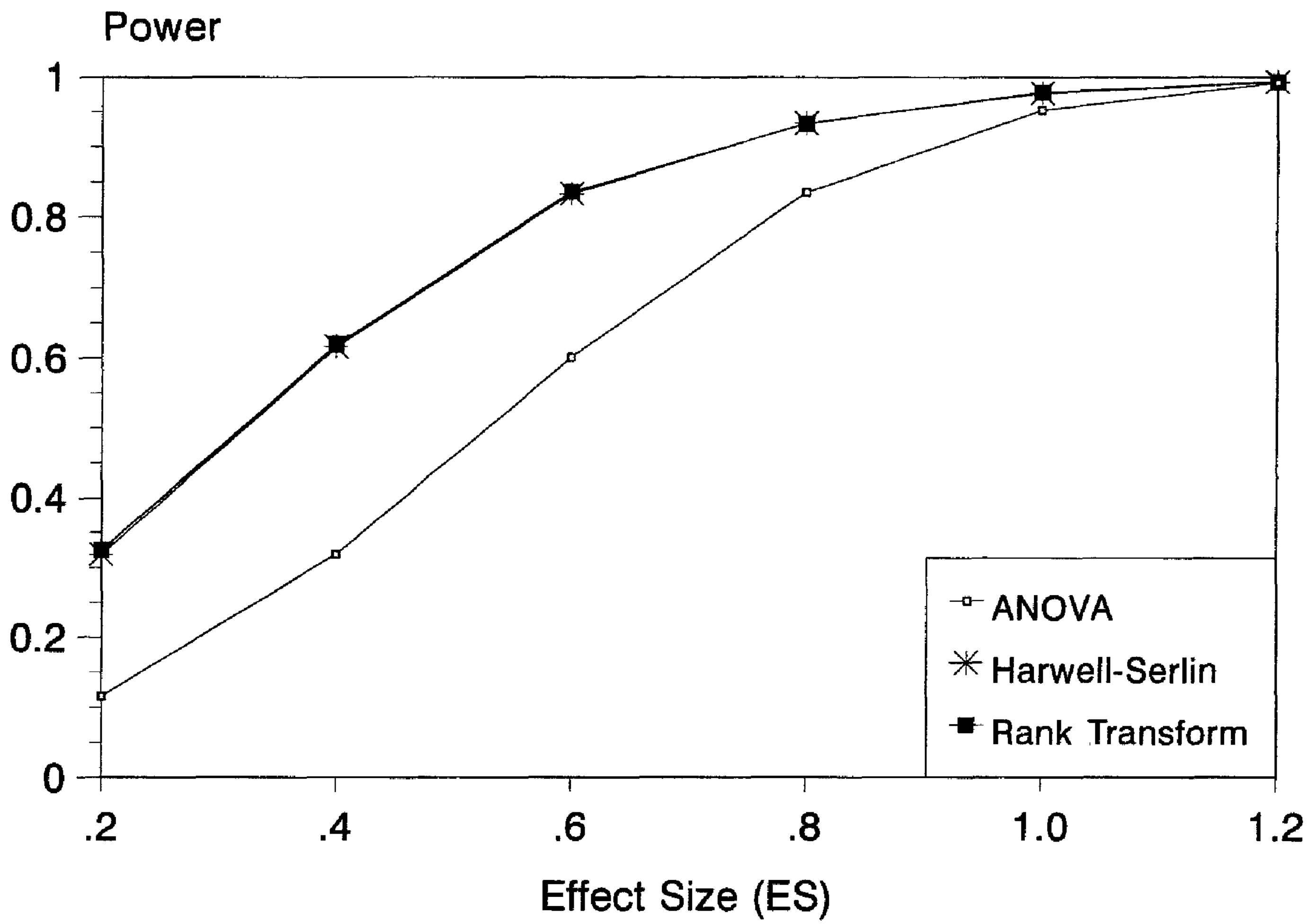

Figure 106. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=7.

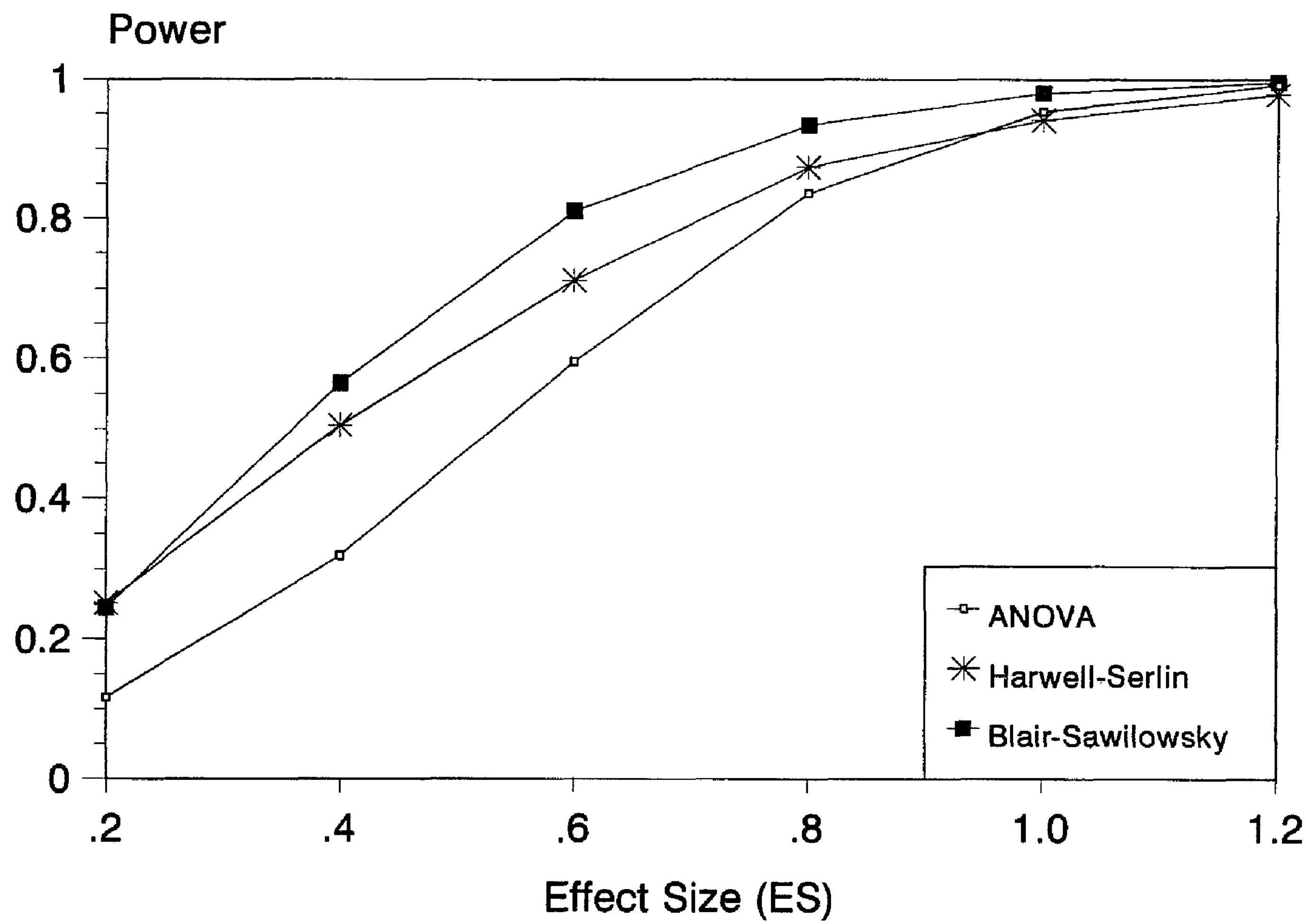

Figure 107. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=7.

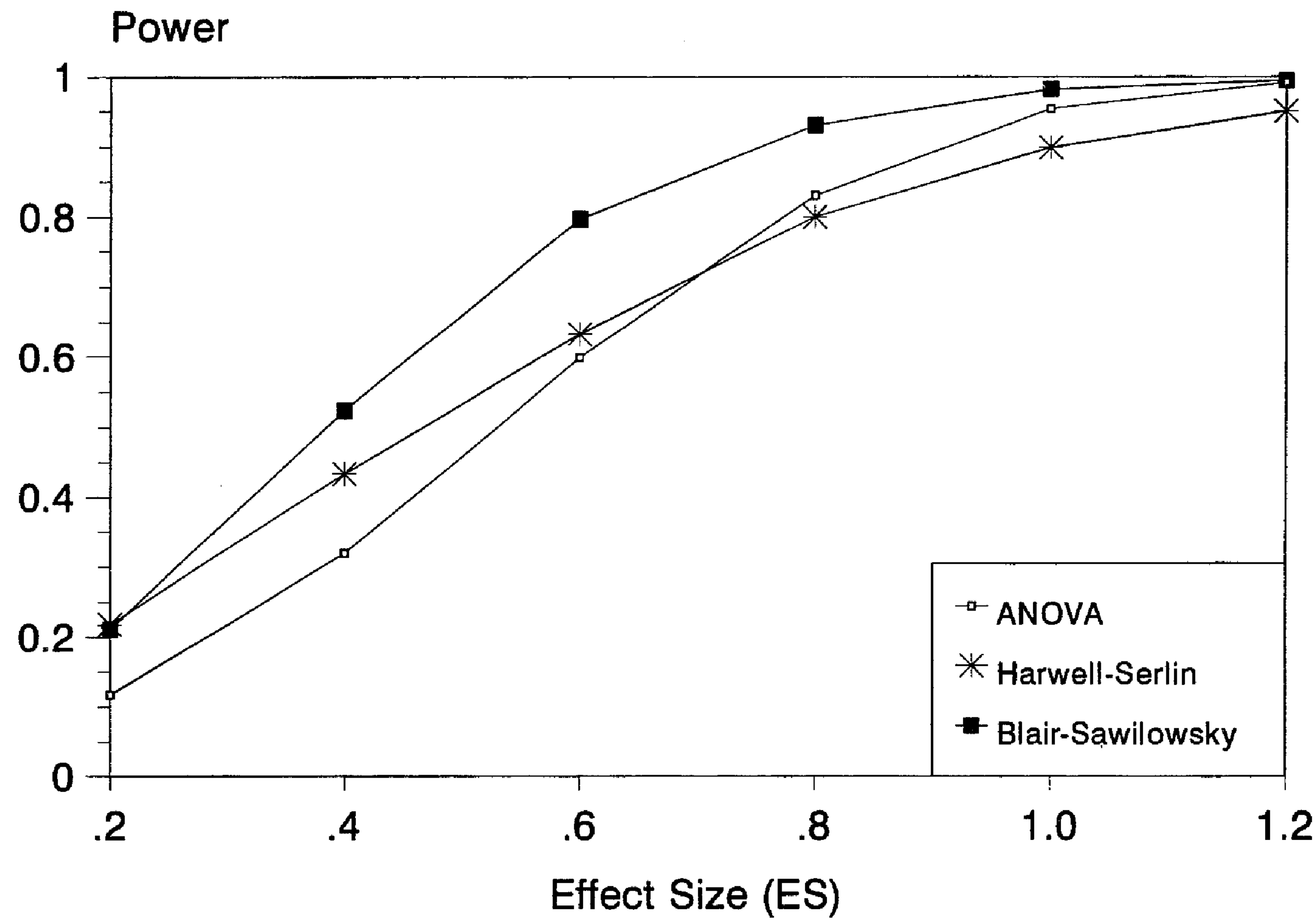

Figure 108. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky when sampling is from the discrete mass at zero data set, alpha=.05 and n=7.

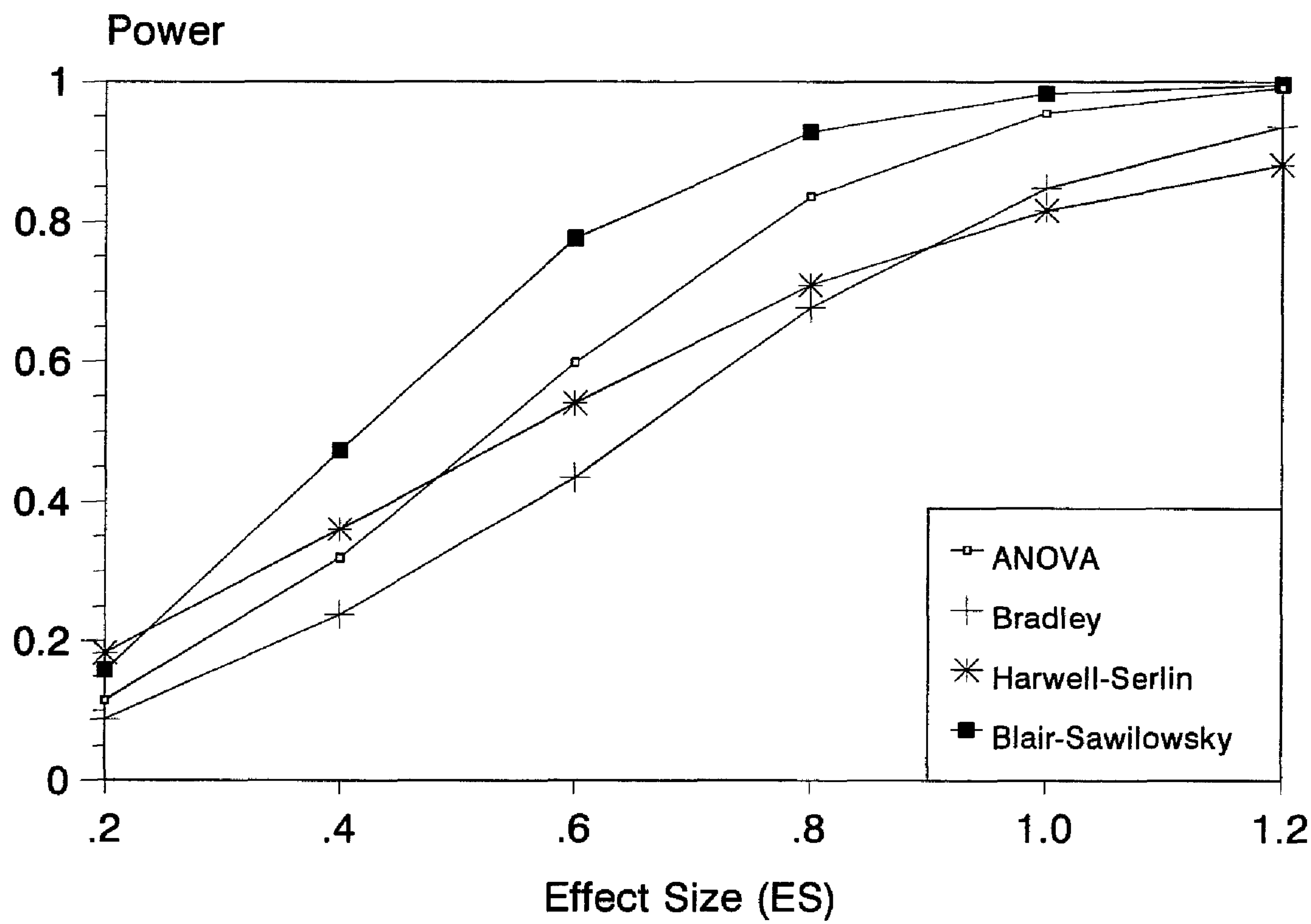

Figure 109. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=7.

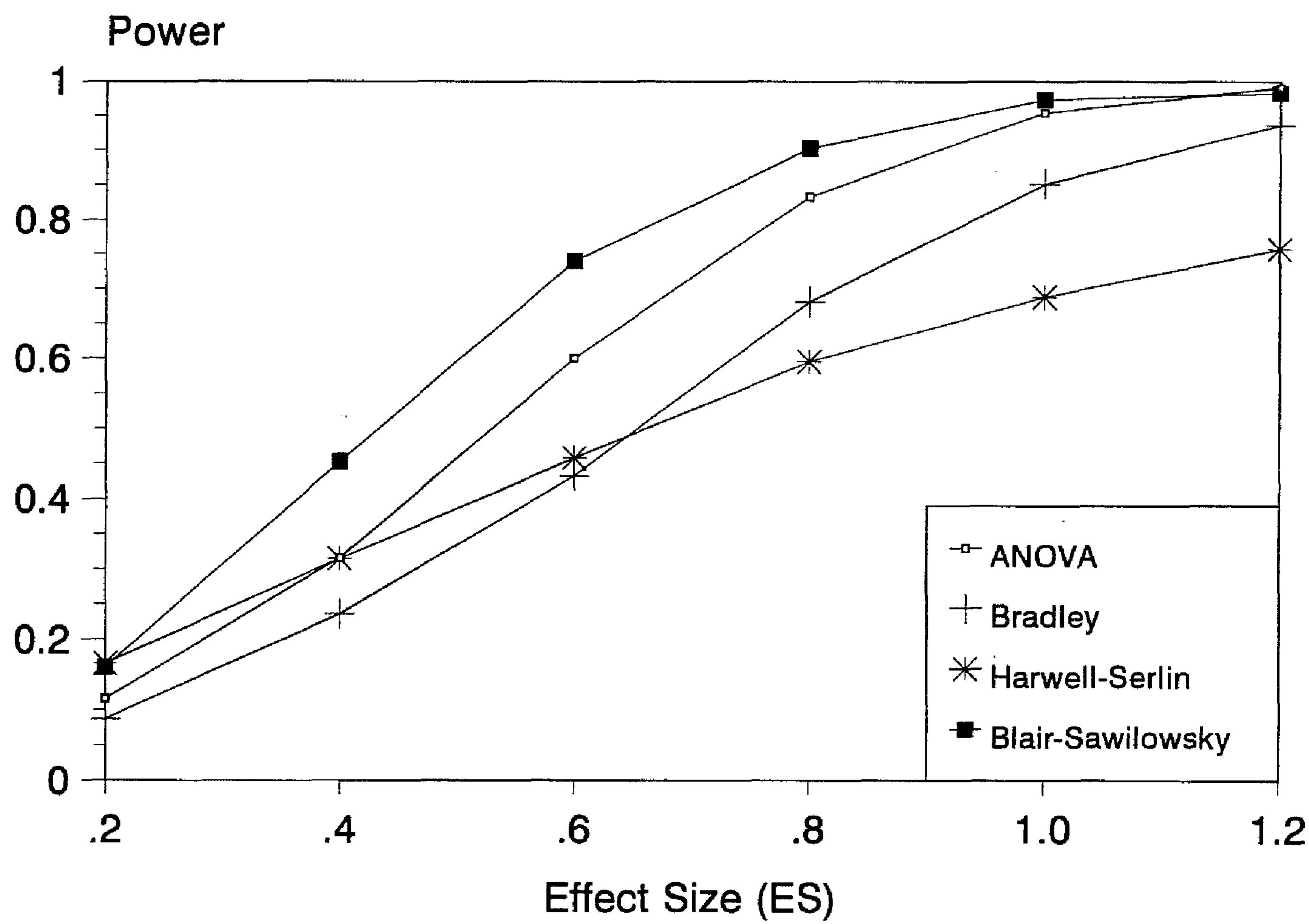

Figure 110. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=7.

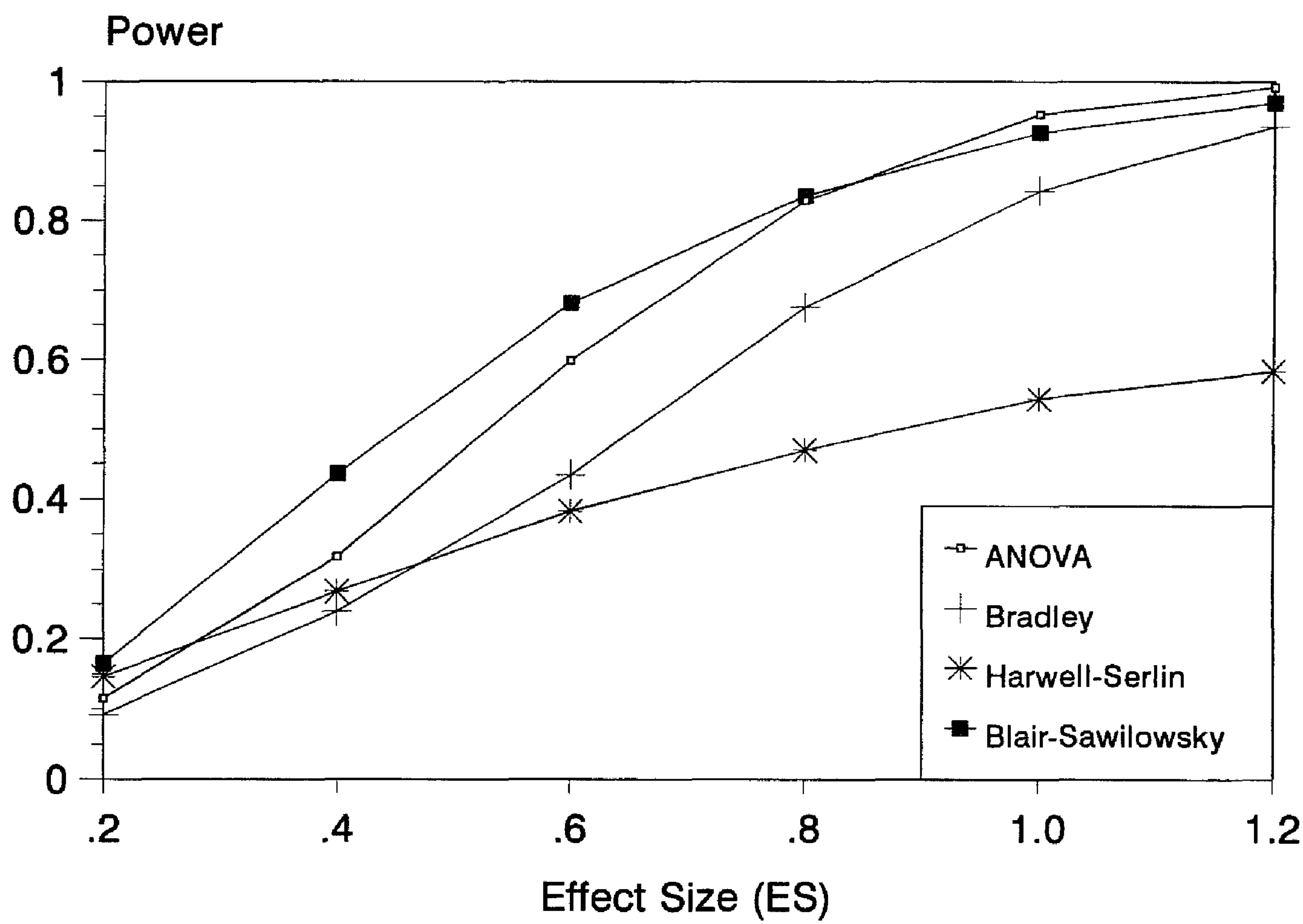

Figure 111. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=7.

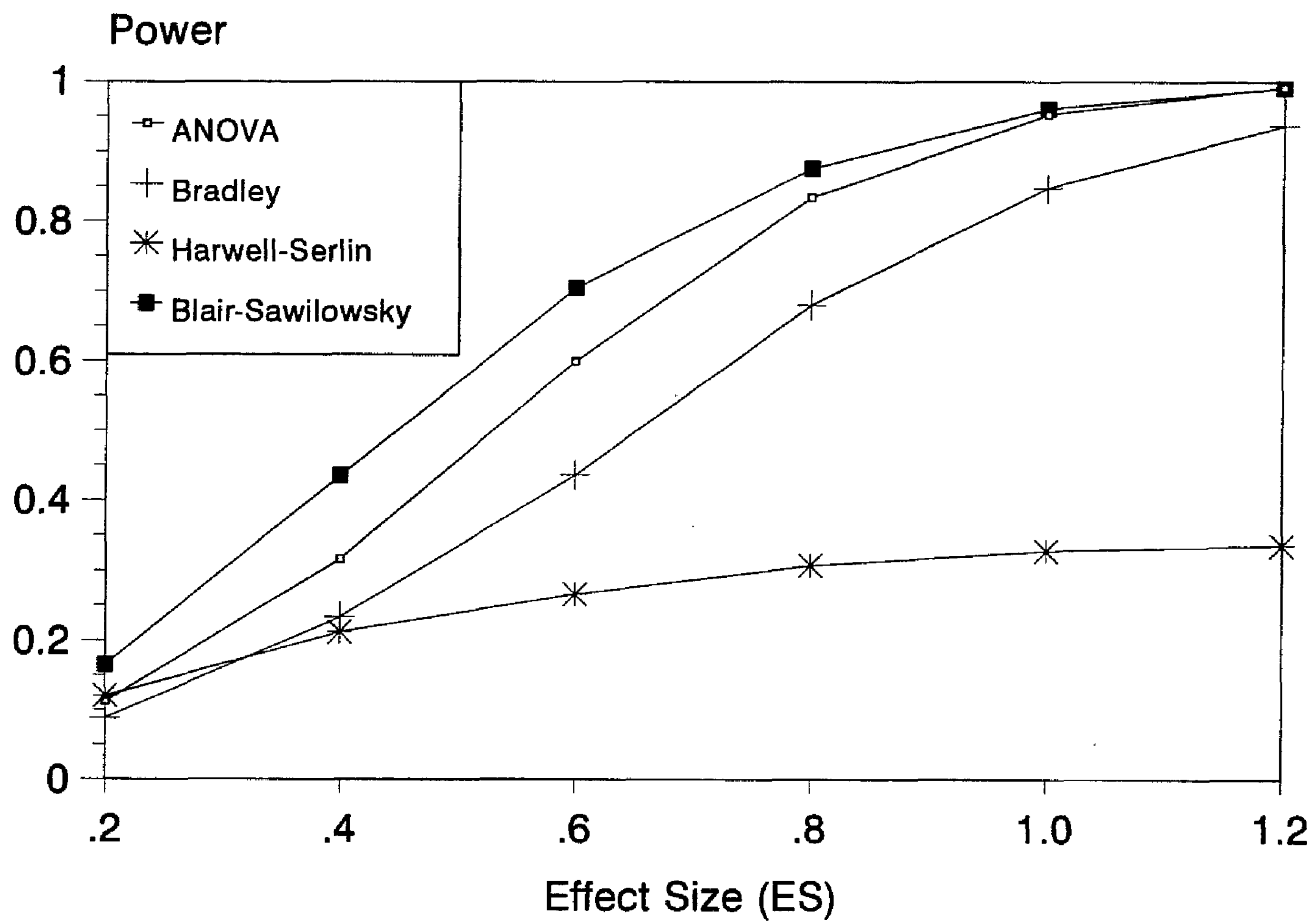

Figure 112. Comparative power of the (abc) interaction effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=7.

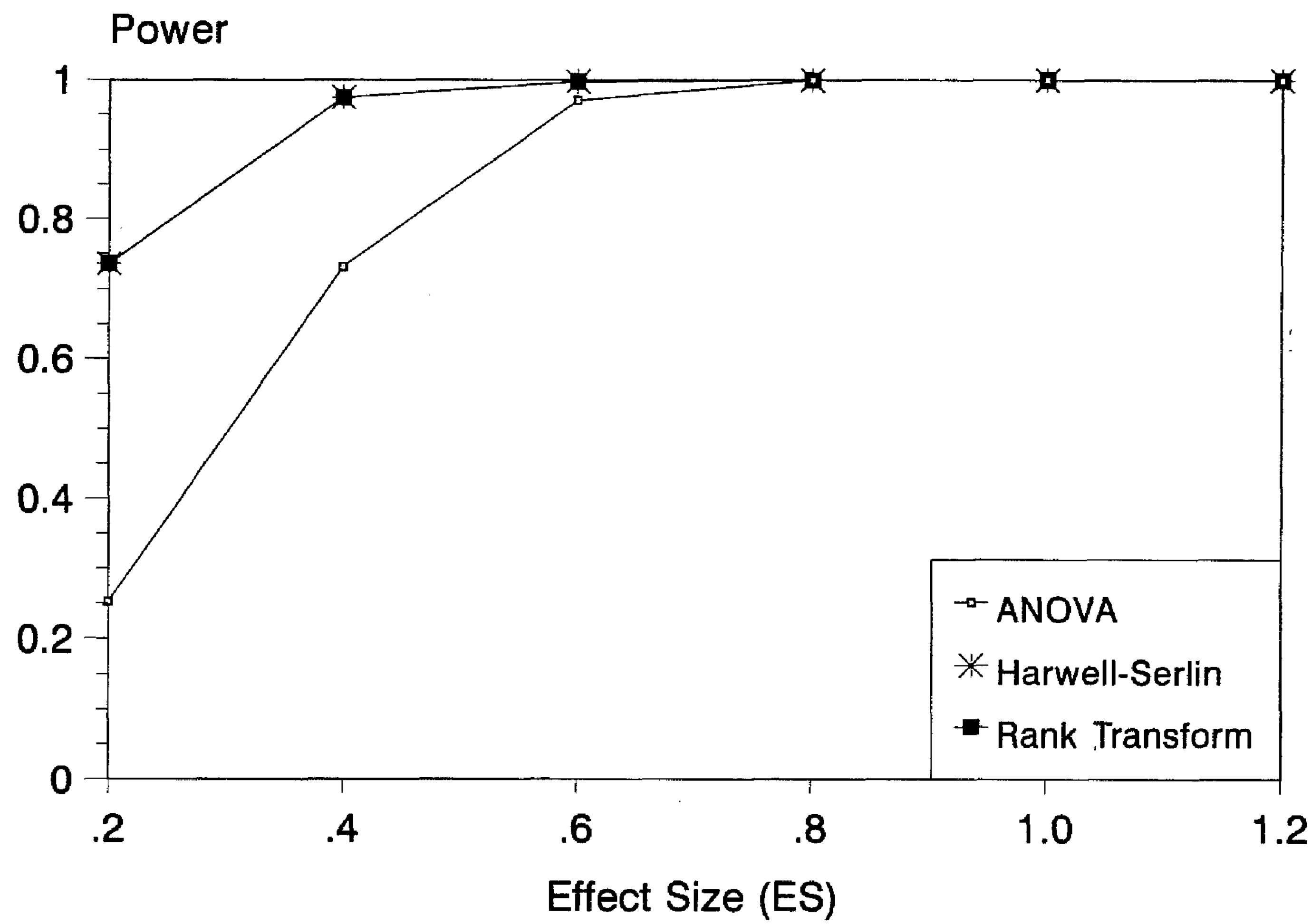

Figure 113. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=21.

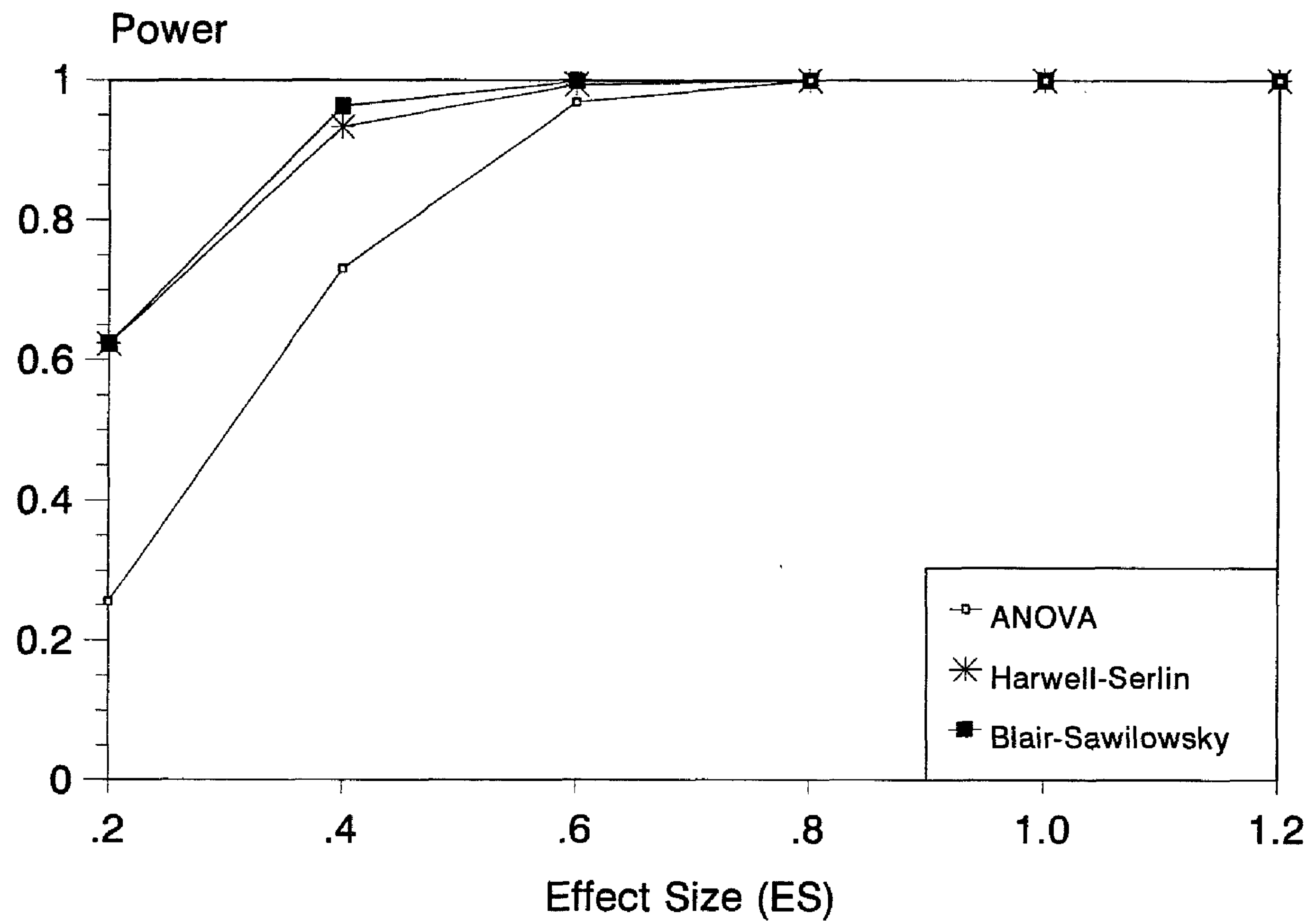

Figure 114. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=21.

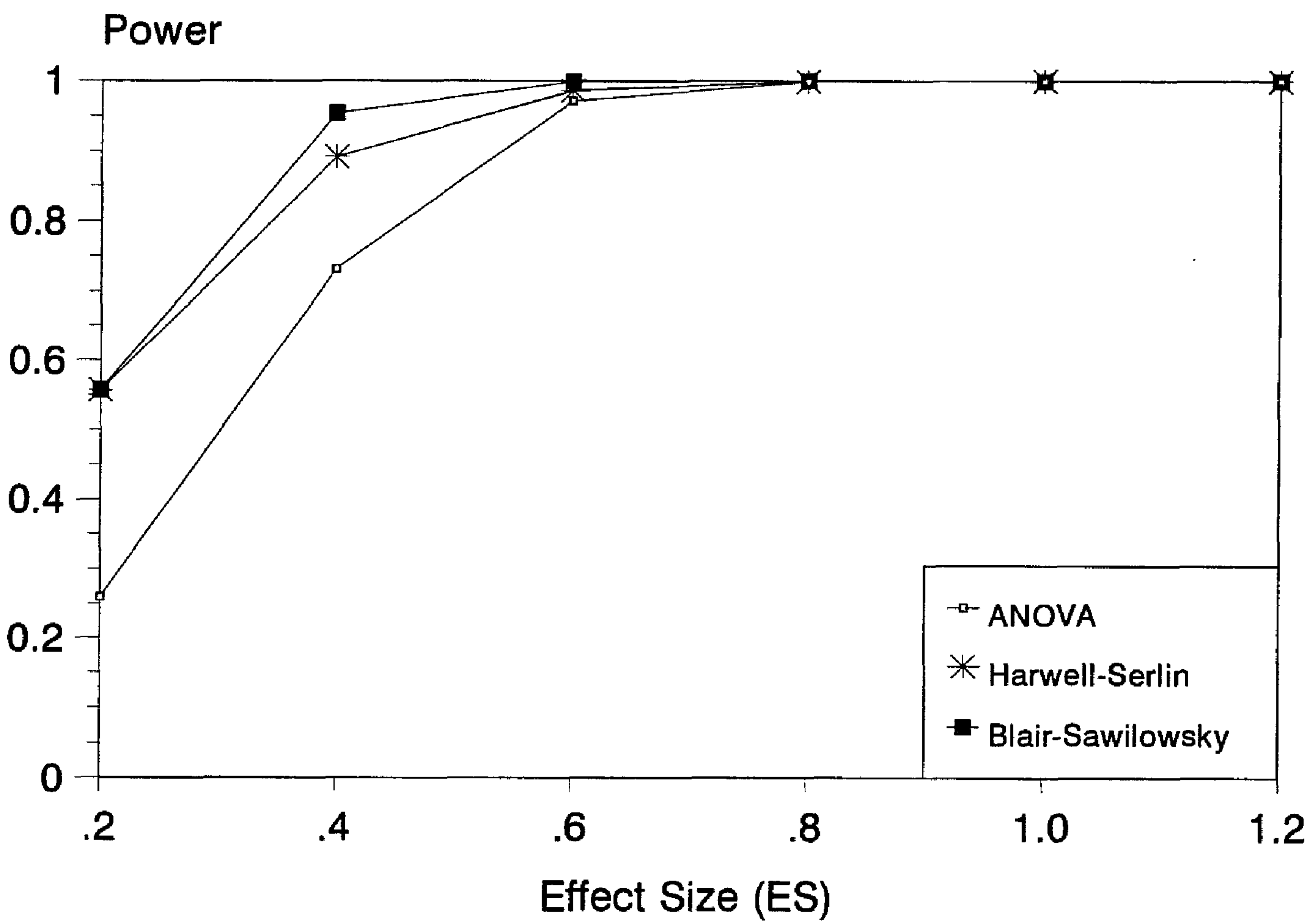

Figure 115. Comparative power of the C main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=21.

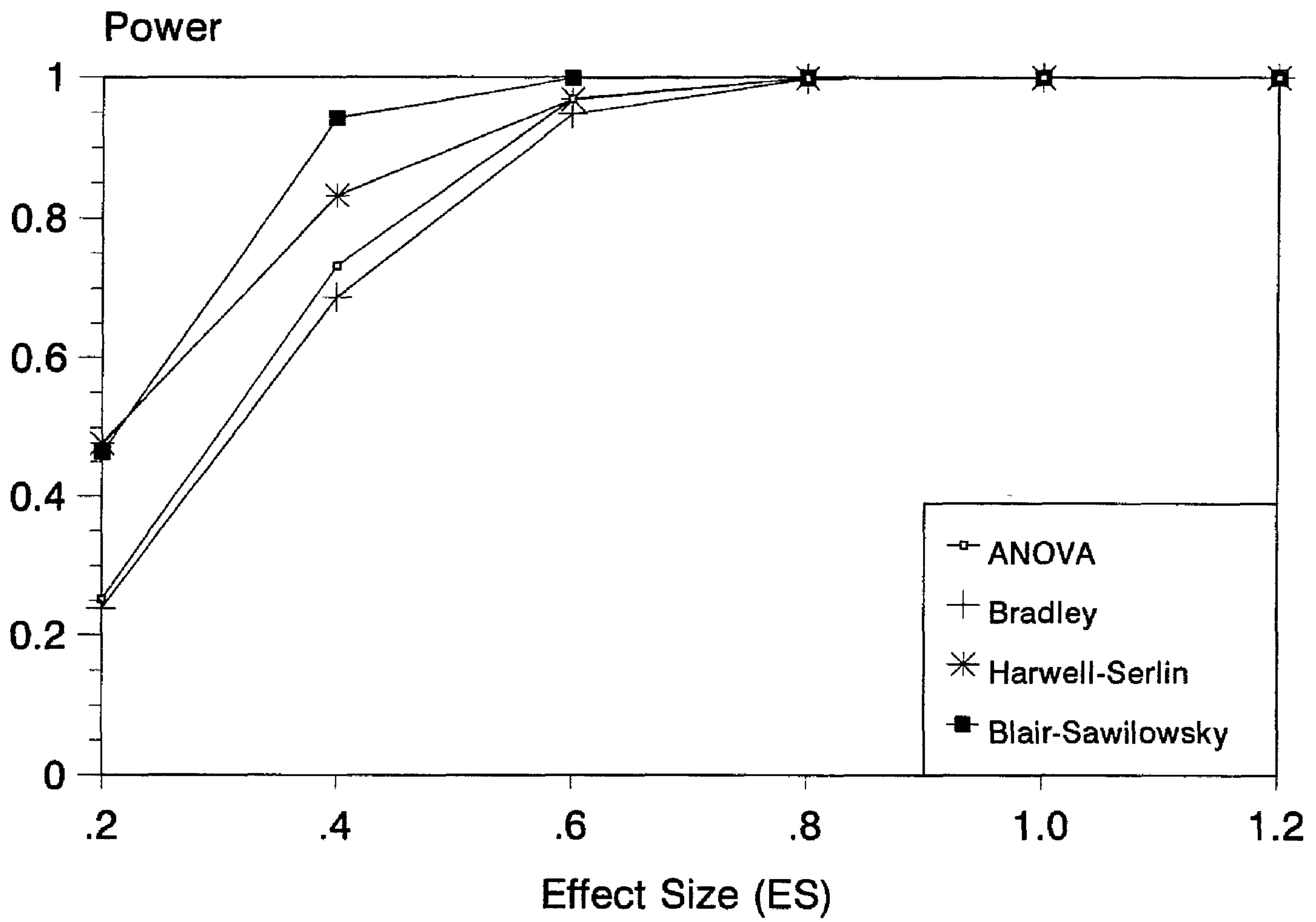

Figure 116. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=21.

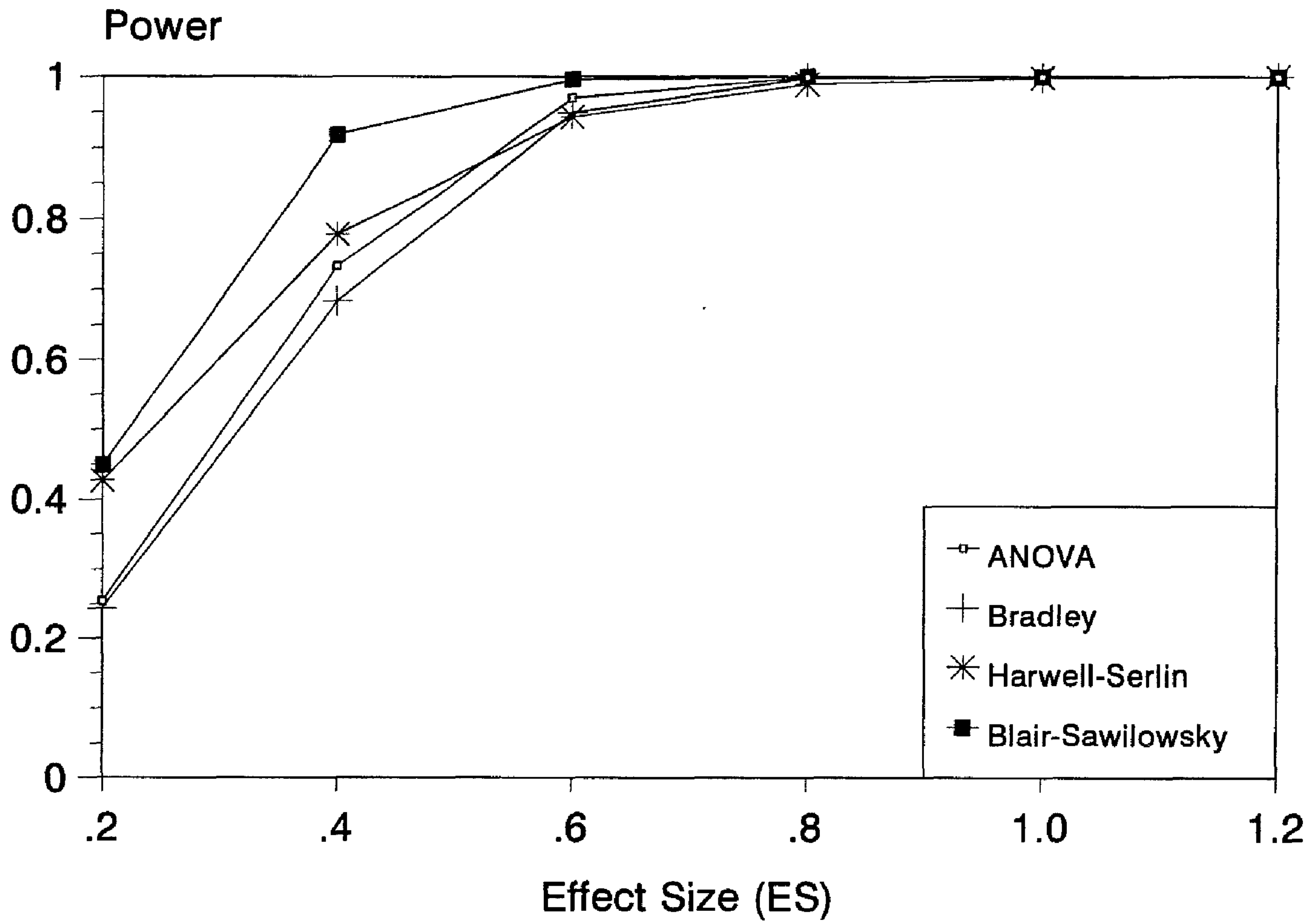

Figure 117. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=21.

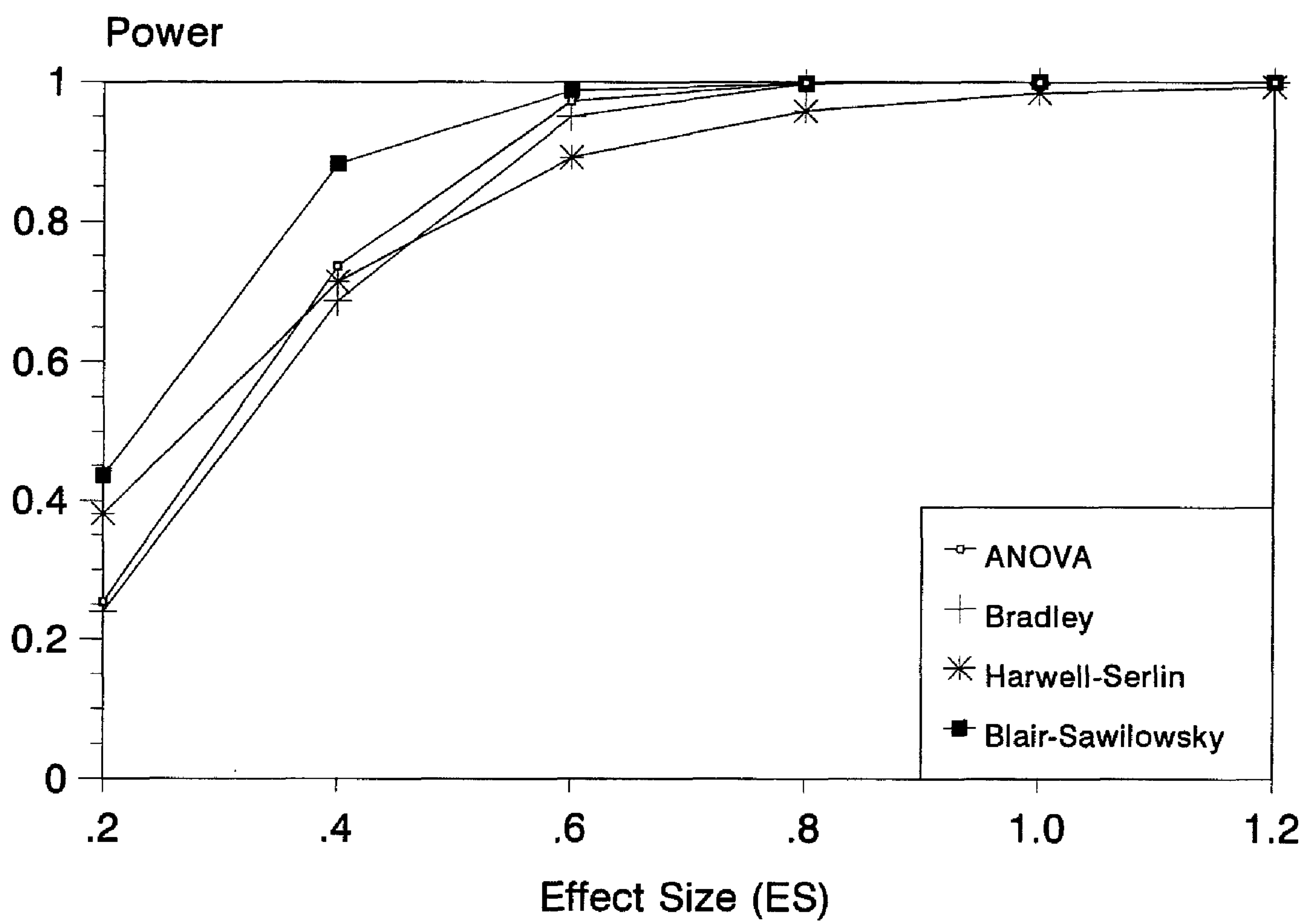

Figure 118. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=21.

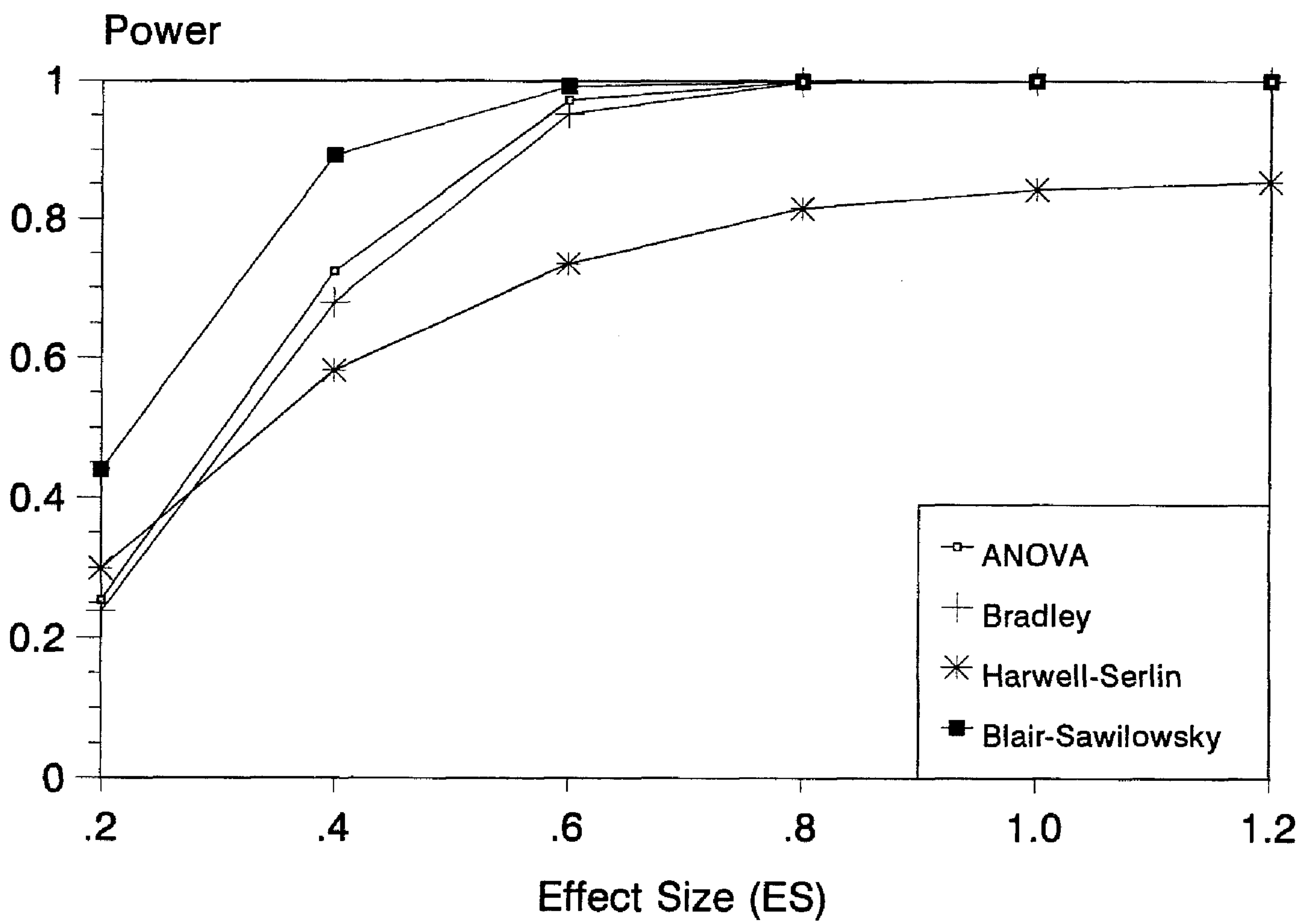

Figure 119. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=21.

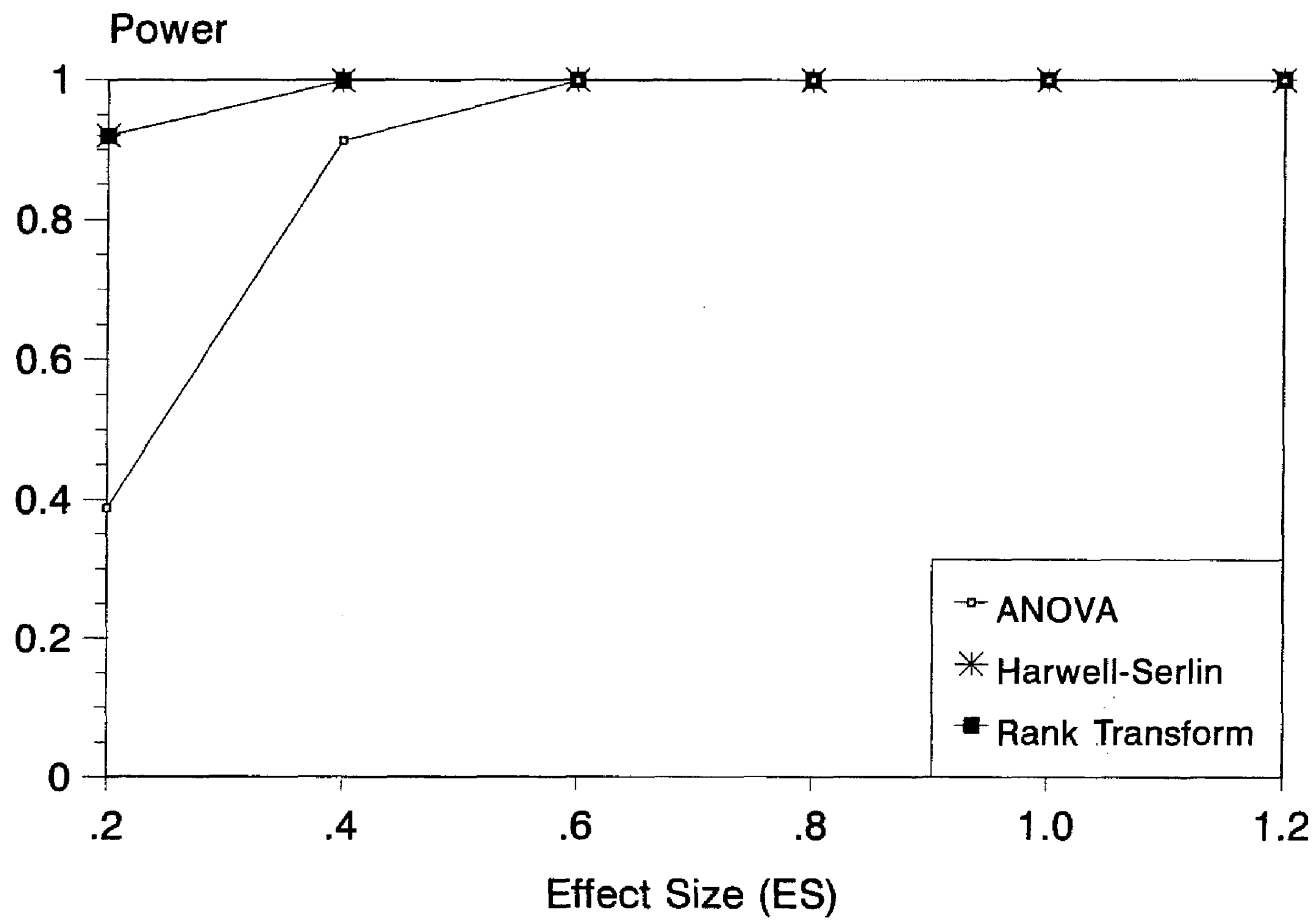

Figure 120. Comparative power of the A main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Rank Transform tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=35.

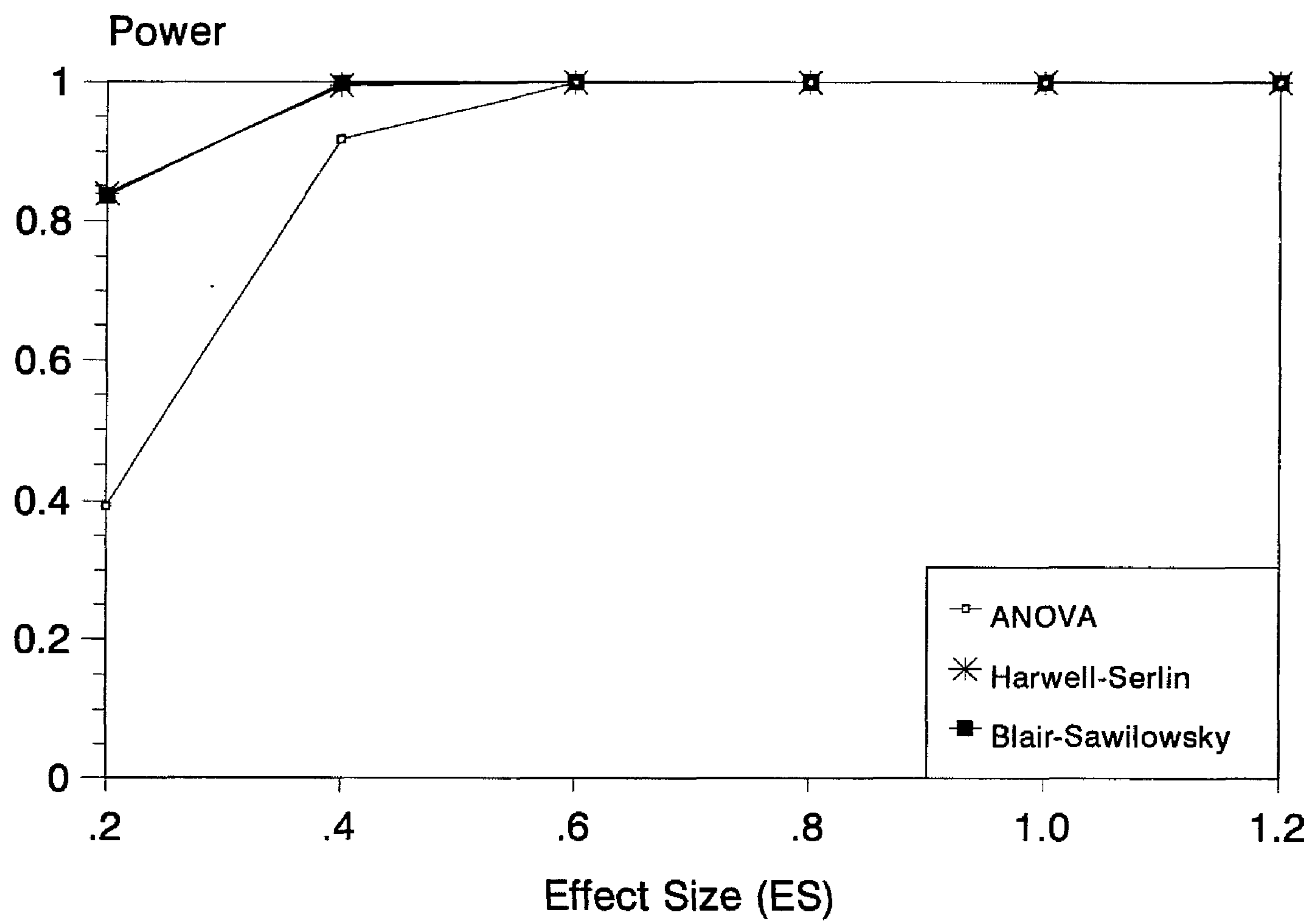

Figure 121. Comparative power of the B main effect for the Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=35.

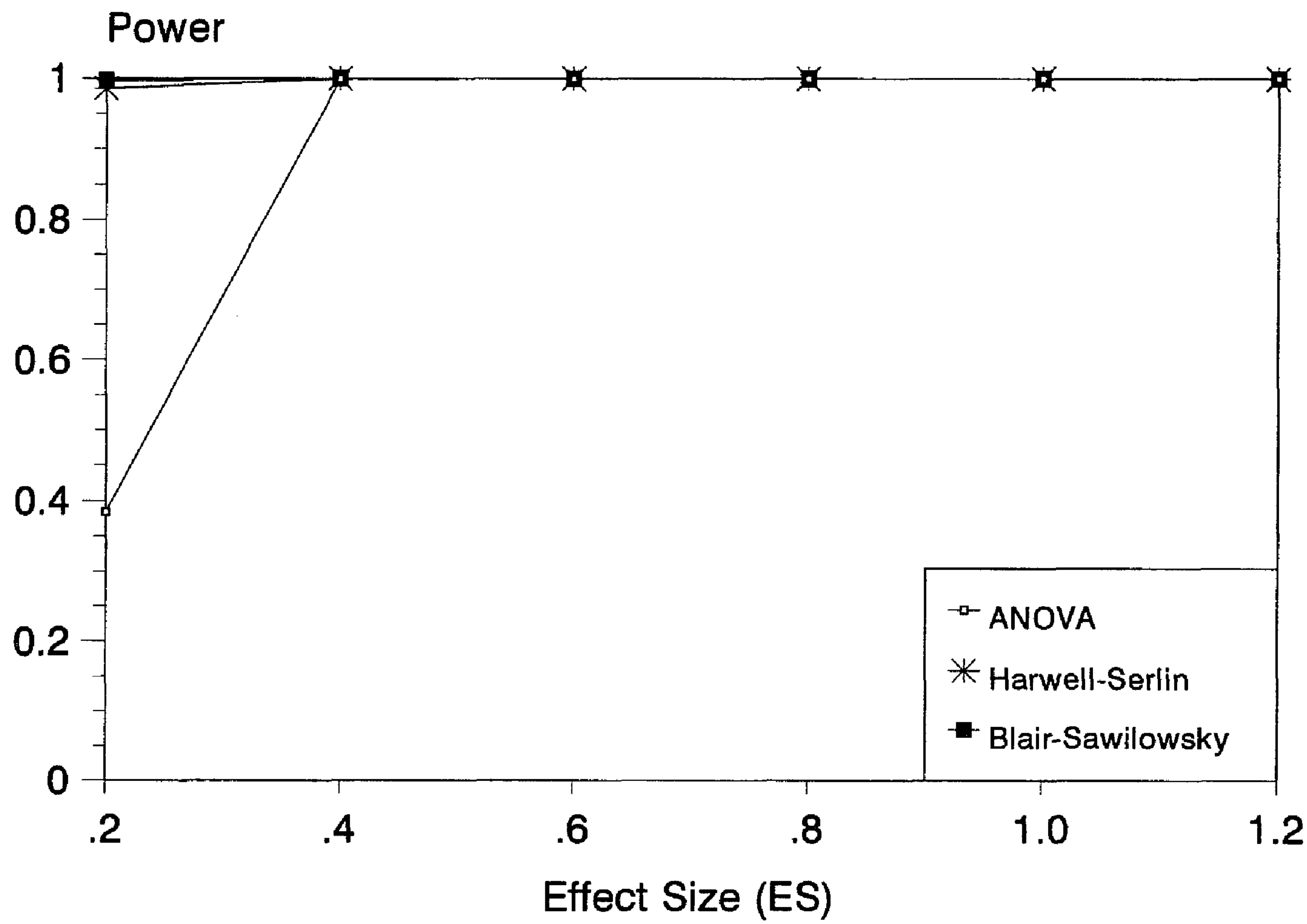

Figure 122. Comparative power of the C main effect for Analysis of Variance (ANOVA), Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=35.

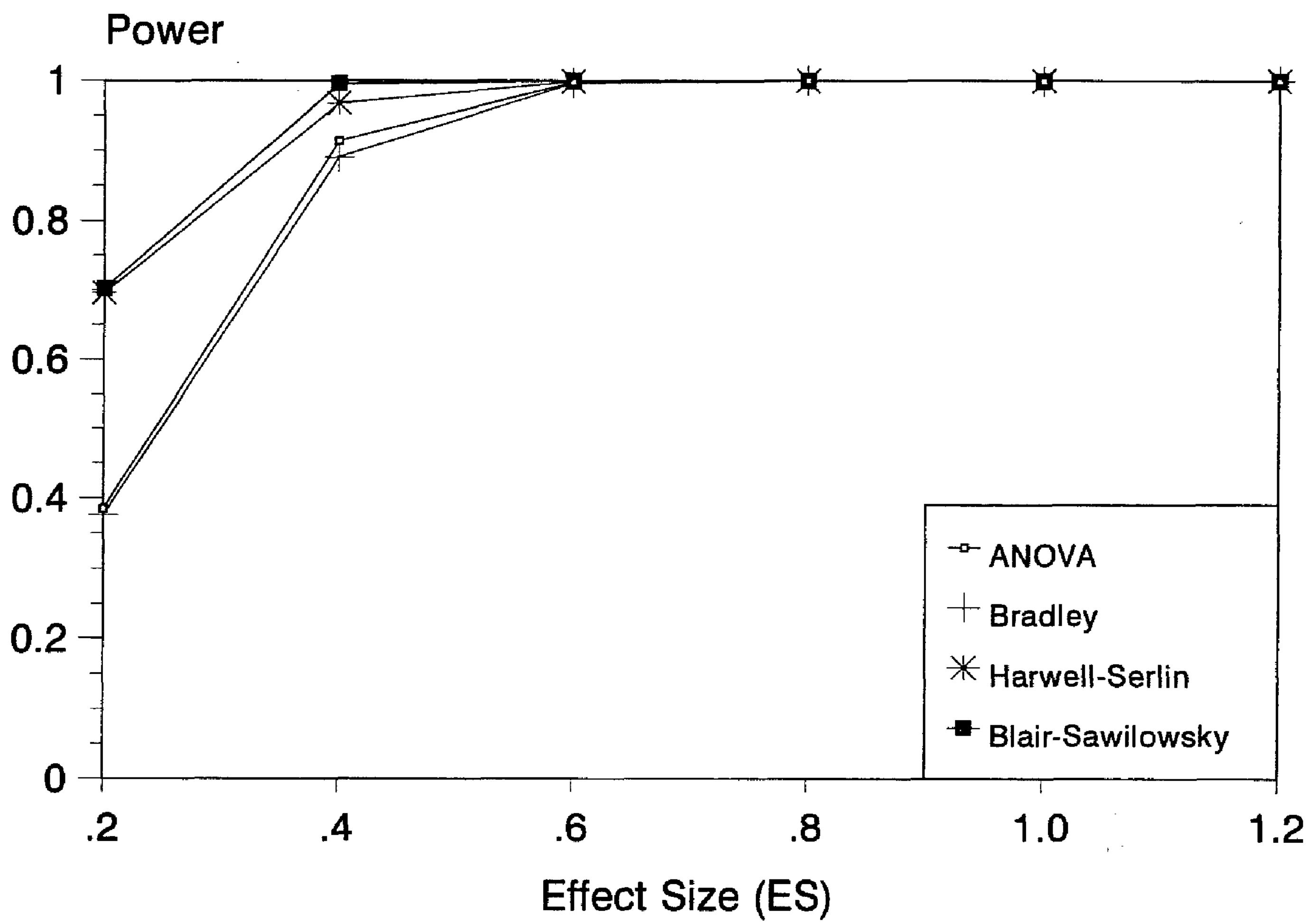

Figure 123. Comparative power of the (bc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=35.

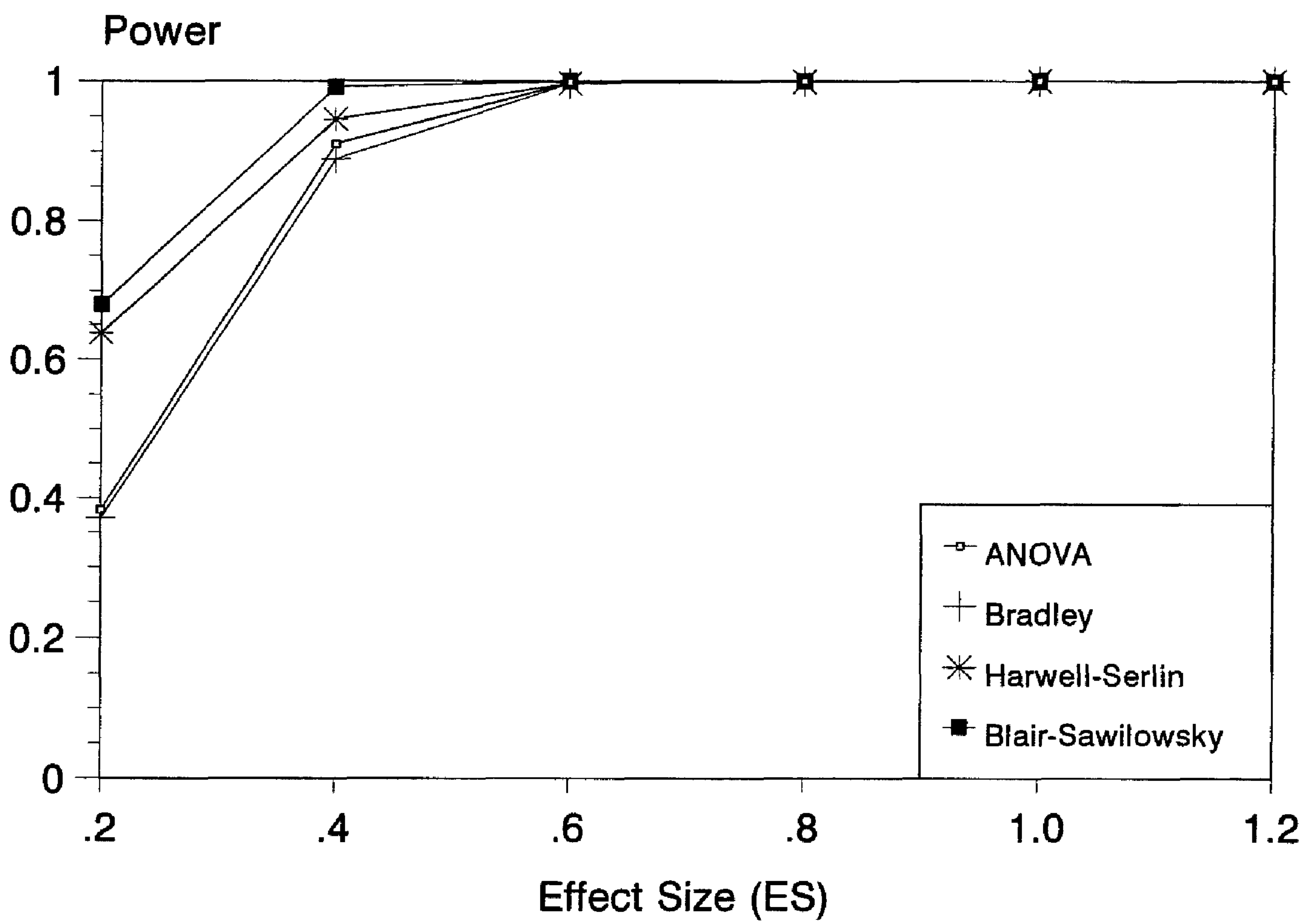

Figure 124. Comparative power of the (ac) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=35.

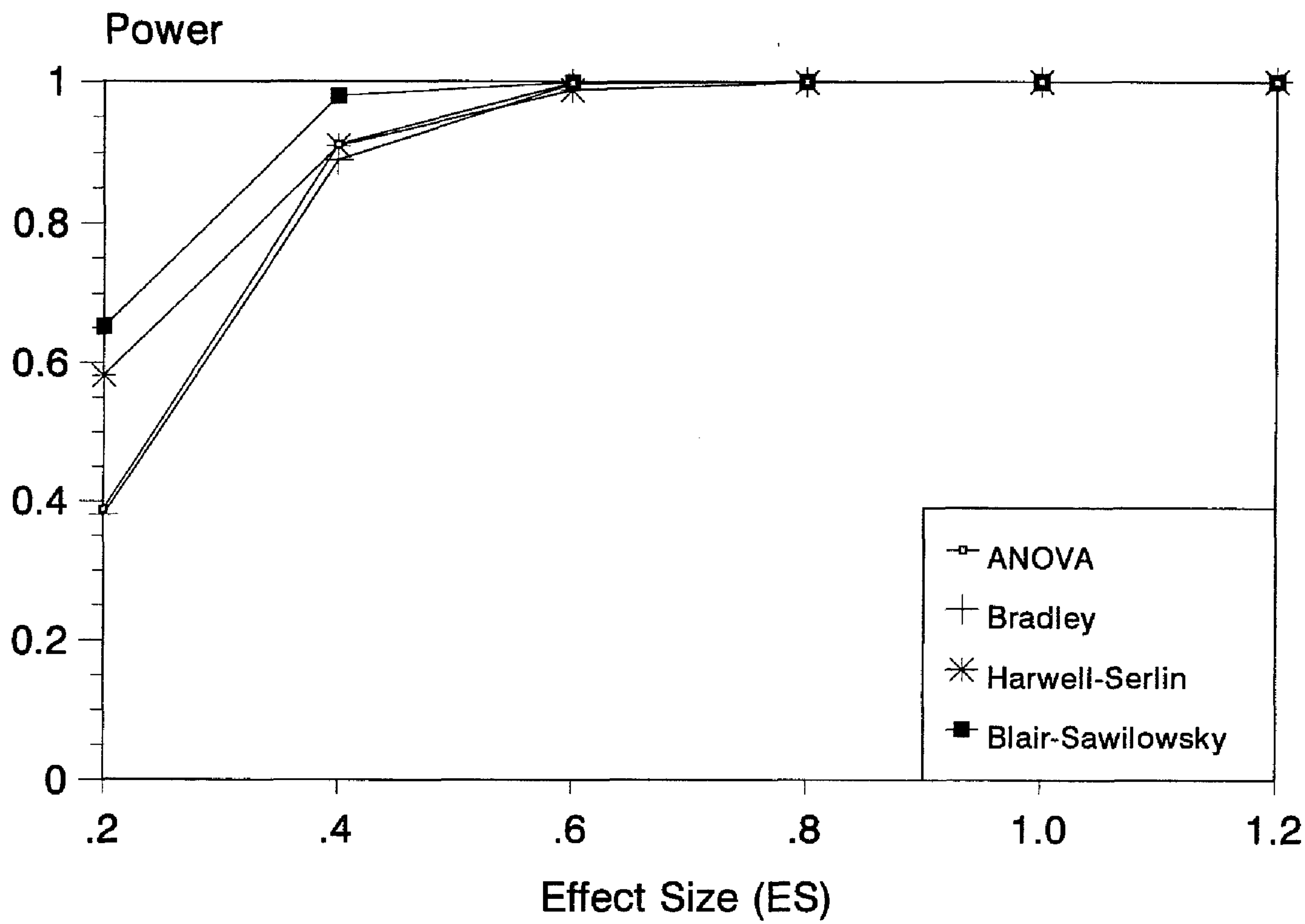

Figure 125. Comparative power of the (ab) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=35.

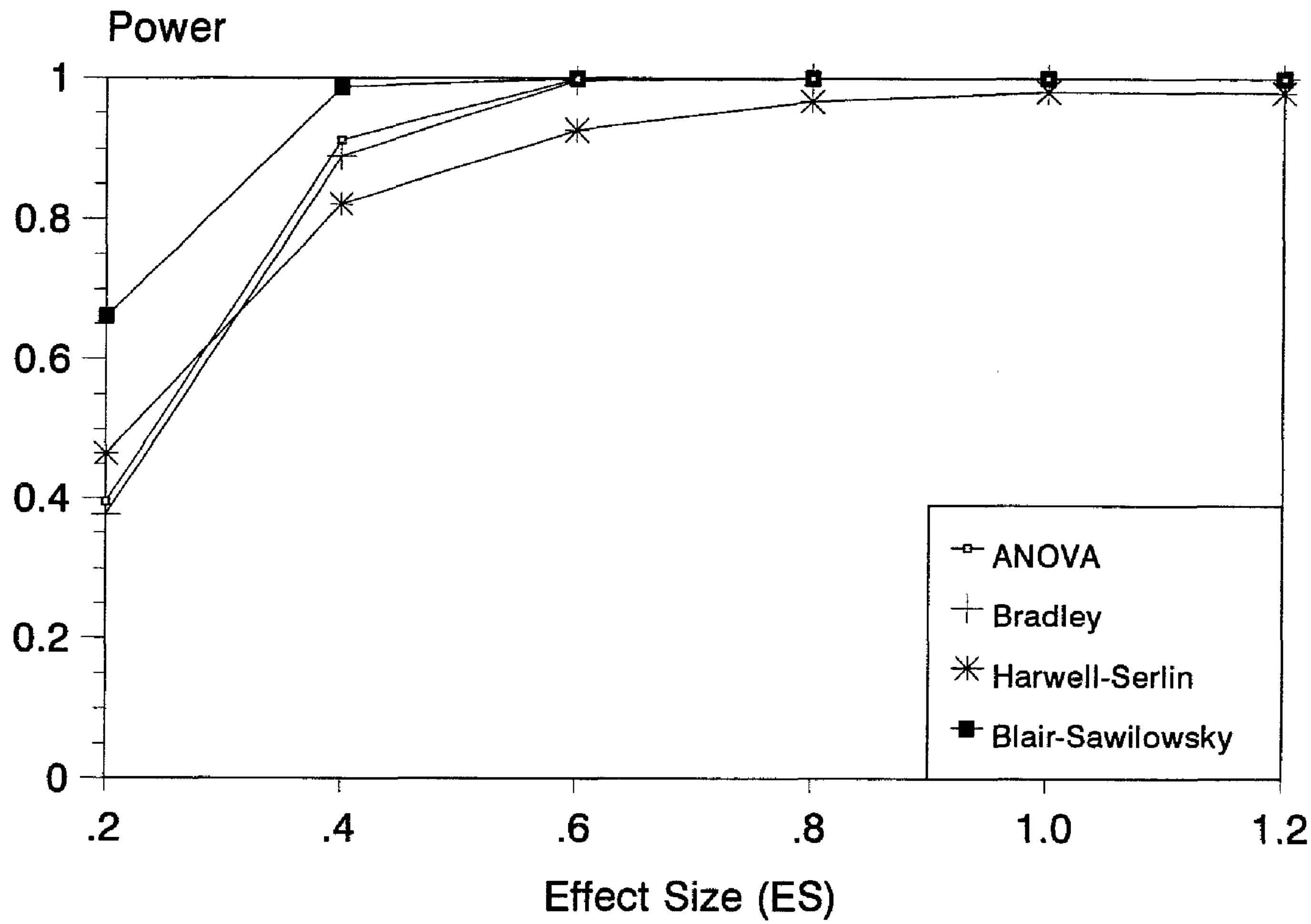

Figure 126. Comparative power of the (abc) interaction for the Analysis of Variance (ANOVA), Bradley, Harwell-Serlin and Blair-Sawilowsky tests when sampling is from the discrete mass at zero data set, alpha=.05 and n=35.

CHAPTER FIVE

CONCLUSIONS AND IMPLICATIONS

Overview of the Study

This study compared the Type I error and power properties of four competing tests: (a) Analysis of Variance, which is a parametric test; (b) Bradley's Collapsed and Reduced Technique, which is a distribution-free test; (c) Harwell and Serlin's L Test, which is a trace criterion form of Puri and Sen's L test (whose substructure is known); and (d) Blair and Sawilowsky's Adjusted Rank Transform test, which is conditionally distribution-free. It is calculationally similar to a test developed independently by Fawcett and Salter (1984). The Fawcett and Salter (1984) version removes the nuisance parameters, as does the Blair-Sawilowsky statistic, but is more computationally involved in that it also subtracts the grand mean.

These four tests were compared under three different sample sizes: (a) n=7, (b) n=21, and (c) n=35. They were also compared under four theoretical distributions and two real data sets: (a) Gaussian distribution, (b) uniform distribution, which represented a light-tailed distribution, (c) t distribution with three degrees of freedom, which represented a heavy-tailed distribution, (d) exponential distribution, which represented a skewed distribution, (e) multi-modal lumpy data set, which is a real education data set of achievement test scores provided by Micceri (1989), and (f) discrete mass at zero data set, which is a real psychology data set of psychometric scores provided by Micceri (1989).

Comparisons were made under eight treatment conditions: (a) all effects were null; (b) all effects were null except for the A main effect which was nonnull; (c) all effects were null except for the A and B main effects which were nonnull; (d) all effects were null except for the A, B and C effects which were nonnull; (e) all effects were null except for the A, B and C main effects and the B x C interaction which was nonnull; (f) all effects were null except for the A, B, and C main effects and the B x C and A x C interactions which were nonnull; g) all effects were null except for the A, B and C main effects and the B x C, A x C, and A x B interactions which were nonnull; h) the A, B and C main effects and the B x C, A x C, A x B and A x B x C interactions were nonnull.

The four tests were compared under six different effect sizes: (a) .2, (b) .4, (c) .6, (d) .8, (e) 1.0, and (f) 1.2. These effect sizes represent the percent of the standard deviation which was added to certain variables in order to model the various treatments.

Restatement of the Purpose of the Study

The purpose of this study is to provide information to individuals in education, psychology and other related disciplines regarding the Type I error rates and power properties of four competing tests. One of the four tests, Analysis of Variance, is usually introduced in basic statistics textbooks with a warning to restrict its use to normally distributed data (Klugh, 1986). The other three tests chosen for this study are distribution-free or conditionally distribution-free, which does not restrict their use to normally distributed data.

This study would be of nominal value if a large proportion of the data used in research was normally distributed. However, in a study conducted by Micceri (1989), of 440 large-sample educational achievement and psychometric measures, it was found that none of the data were normally distributed. Micceri (1989) concluded that test scores are rarely normally distributed.

The Kruskal-Wallis and Friedman tests are widely accepted as distribution-free alternatives to ANOVA (Ingram & Monks, 1992) when there is no interaction in the data set. However, it has been noted that interaction effects are fairly common (Tate & Clelland, 1957) and there is a lack of nonparametric tests for interaction (Anderson, 1961; Bradley, 1968; Toothaker & Chang, 1980).

Sawilowsky (1990) presented ten non-parametric tests for interaction, which are alternatives to ANOVA and have been developed over the past 25 years. The majority of the tests presented by Sawilowsky were developed during the 1980s. This study specifically chose three of the ten alternatives for examination because all three met the following criteria: (a) may be easily calculated, (b) may be performed with existing computer packages, and (c) appeared to have the most potential as competitors to the parametric ANOVA test when the underlying assumption of normality is violated.

This study uses Bradley's (1978) liberal criterion for robustness. For a nominal .05 alpha level, the acceptable actual Type I error rate for robustness ranges from .025 to .075.

All the comments in this chapter are restricted to the hypothesis associated with shift in location parameter. In addition, the discussion in this paper applies to a 2 x 2 x 2 layout only.

Adequacy of Algorithms and Type I error

Inspection of Table 11 when all effects are null indicates that the ANOVA, Bradley and Harwell-Serlin were within sampling error of nominal alpha under the

Gaussian distribution. This demonstrates the adequacy of the algorithms used in the Fortran computer program. The Type I error results for the remaining distributions, when all effects were null, indicate that the procedures were within sampling error of nominal alpha. This means that the parametric and conditionally distribution-free tests were robust for these distributions.

Tables 12 and 13

Tables 12 and 13 show the Type I error and power results when the number of nonnull effects increased from 1 to 7 under the Gaussian distribution. Table 12 indicates the results when sample size was $\underline{n}=7$, and Table 13 indicates the results when sample size was $\underline{n}=35$. Because the Analysis of Variance F test is by definition the Uniformly Most Powerful Unbiased (UMPU) test under normality, it was the most powerful test for all conditions studied under this population. Blair-Sawilowsky's power was only slightly less powerful than the ANOVA in these two tables. For example, ANOVA's power ranged from .113-.115 and Blair-Sawilowsky's power ranged from .111-.114 when $\underline{c}=.2$ and $\underline{n}=7$. (Recall that the averages recorded under Column 1 for Blair-Sawilowsky reflect the unadjusted RT only, and not the Adjusted Rank Transform, as discussed in Chapter Four.) The trend continued throughout the tables, with Blair-Sawilowsky's power trailing slightly behind the ANOVA. Initially, Harwell-Serlin was third in order of power, and Bradley's test was least competitive. However, as the effect and sample size increased, and as more treatments were introduced, Harwell-Serlin's performance decreased, becoming less powerful than Bradley.

The second aspect of the Tables 12 and 13 are the Type I error rates for null effects. According to Bradley's liberal definition of robustness, ANOVA, Bradley and

Blair-Sawilowsky remained robust throughout the tables. Harwell-Serlin was robust when the effect size was small. However, as the effect size increased, Harwell-Serlin became non-robust in the conservative manner (i.e., alpha was less than .025).

Tables 14 and 15

Tables 14 and 15 show the Type I and power results under the uniform distribution when sample size was $\underline{n}=7$ (Table 14) and $\underline{n}=35$ (Table 15). The most powerful test under these conditions was the ANOVA. Blair-Sawilowsky was slightly less powerful than ANOVA. When the effect size was small ($\underline{c}=.2$) and the sample size was small ($\underline{n}=7$), Harwell-Serlin was more powerful than Bradley (but less than Blair-Sawilowsky). However, as indicated on the tables under the other conditions ($\underline{n}=35$ or $\underline{c}=.8$), Harwell-Serlin became less powerful than Bradley. Harwell-Serlin also showed a tendency to decrease in power as the number of nonnull effects increased. For example, Table 15 indicates when $\underline{n}=35$ and $\underline{c}=.8$, Harwell-Serlin started off with a power level greater than .9999 under the condition of one non-null effect. When the condition of seven nonnull effects was reached, the power dropped to .979.

The Type I error rates remained consistently robust for all tests when $\underline{c}=.2$. However, when $\underline{c}=.8$, Harwell-Serlin showed conservative nonrobustness as the alpha levels dropped below .025.

Tables 16 and 17

Tables 16 and 17 show the Type I and power results for the t distribution with three degrees of freedom when the sample size was $\underline{n}=7$ (Table 16) and when sample size was $\underline{n}=35$ (Table 17). When the effect size was small ($\underline{c}=.2$), the tables show Blair-Sawilowsky as the most powerful test and Harwell-Serlin as the second most powerful. Although Blair-Sawilowsky's power remained consistent across treatments, Harwell-Serlin's power steadily decreased across treatments indicating a difficulty in detecting a false null hypothesis when the number of treatments increased. Bradley ranked third in power, when $\underline{n}=7$, and ANOVA ranked last; however, when $\underline{n}=35$, these two switched places. The power advantage of the RT over ANOVA when $\underline{c}=.2$ and $\underline{n}=35$ was quite large. For example, Blair-Sawilowsky's power was .631 and ANOVA showed a power level of only .431 when there were two nonnull effects.

When the effect size increased ($\underline{c}=.8$) at $\underline{n}=7$, Blair-Sawilowsky remained the most powerful. Harwell-Serlin initially had a similar power level, but quickly lost power as the number of nonnull effects increased, and dropped to last place. ANOVA moved to second place and became more powerful than both Bradley and Harwell-Serlin.

The Type I error rates for all of the tests except Harwell-Serlin were robust. When the effect size increased ($\underline{c}=.8$) Harwell-Serlin initially experienced deflated alpha levels of .013 and .008. As the number of nonnull effects increased, Harwell-Serlin's Type I error rose dramatically to an inflated value of .131. Because these levels are less than .025 and greater than .075 Harwell-Serlin is considered nonrobust according to Bradley's liberal definition. Also, it should be noted that the impact of conservative results are reduced power, whereas liberal results are more severe and are evidence of a test being invalid.

Tables 18 and 19

Tables 18 and 19 show the Type I and power results when sampling was from the exponential distribution, $\underline{c}=.2$ and .8, and sample size was $\underline{n}=7$ (Table 18) and $\underline{n}=35$ (Table 19). The most powerful test under the conditions outlined in Tables 18 and

19 was Blair-Sawilowsky. Harwell-Serlin started at approximately the same level, but as the number of effects increased the power quickly dropped. When the sample size and effect size were small ($\underline{n}$=7, $\underline{c}$=.2) Harwell-Serlin remained the second-most powerful test, even with the drop. However, when the effect size increased ($\underline{c}$=.8), Harwell-Serlin became the least competitive test. ANOVA ranked third when Harwell-Serlin retained second place; however, when Harwell-Serlin dropped, ANOVA moved to second place. Many times, the power level of Blair-Sawilowsky was almost twice that of the ANOVA. For example, it was .200 to .118 when $\underline{n}$=7 and $\underline{c}$=.2 and there were two nonnull effects. Bradley followed ANOVA in third place when the sample size was $\underline{n}$=35, and last place when sample size was $\underline{n}$=7.

As in the other tables, all of the tests were robust except for Harwell-Serlin which was both conservatively and liberally nonrobust when the effect size increased ($\underline{c}$=.8). In fact, the Type I error increased to 0.147 when there were six non-null effects.

Tables 20 and 21

Tables 20 and 21 show the Type I error and power results when the number of nonnull effects increased from 1 to 7 under the multi-modal lumpy data set. Table 20 indicates the results when sample size was $\underline{n}$=7, and Table 21 indicates the results when sample size was $\underline{n}$=35.

When the effect size was small ($\underline{c}$=.2) the most powerful test was Blair-Sawilowsky. ANOVA was slightly less powerful under this effect size. However, when the effect size increased to $\underline{c}$=.8 and $\underline{n}$=7, ANOVA became slightly more powerful than Blair-Sawilowsky. Harwell-Serlin was initially almost as powerful as Blair-Sawilowsky and more powerful than ANOVA under all conditions when there was only one or were only two nonnull effects. However, as the number of nonnull effects increased, Harwell-Serlin lost power and fell to last or second-to-last place. Bradley exhibited power levels below both ANOVA and Blair-Sawilowsky in all cases except when the sample size was large ($\underline{n}$=35) and the treatment was large ($\underline{c}$=.8). Under these conditions, all tests exhibited an initial power greater than .9999. All tests retained their high power except Harwell-Serlin, which dropped to .977 when the number of nonnull effects were 7.

All of the tests were robust except for Harwell-Serlin, which was both conservatively and liberally nonrobust when the power level increased to $\underline{c}$=.8.

Table 22 and 23

Tables 22 and 23 show the Type I error and power results when the number of nonnull effects increased from 1 to 7 under the discrete mass at zero data set. Table 22 indicates the results when sample size was $\underline{n}$=7, and Table 23 indicates the results when sample size was $\underline{n}$=35.

Blair-Sawilowsky was the most powerful test under these conditions, with power levels more than double ANOVA in several cases. When $\underline{n}$=7, $\underline{c}$=.2 and the number of nonnull conditions was 2, Blair-Sawilowsky had a power level of .245 and ANOVA had a power level of .117. When there were few nonnull conditions, Harwell-Serlin was approximately the same or slightly less powerful than the Rank Transform. However, as the number of nonnull effects increased, Harwell-Serlin exhibited at least a 50% decrease in power, and as much as a 67% decrease, in all but the condition when $\underline{n}$=35 and $\underline{c}$=.8. For example, when $\underline{c}$=.8 and $\underline{n}$=7, Harwell Serlin went from a power of .934 to .312. ANOVA was more competitive than Bradley, which was the least competitive test when c=.2, and the third most powerful when c=.8. Harwell was more competitive than ANOVA when c=.2; however, when c=.8, Harwell-Serlin dropped to last place.

All of the tests were robust when c=.2. However, when c=.8, ANOVA and Bradley Type I errors remained at approximately .05, while Harwell-Serlin both inflated and deflated to become conservatively and liberally non-robust. Harwell-Serlin had a conservative Type I error rate of .019, and a liberal Type I error rate of .739 when n=35 and c=.8. Blair-Sawilowsky Type I error increased slightly to .061 when n=7 and c=.8, and .075 when n=35 and c=.8, but remained robust.

Figures 1 through 126

These figures are useful in showing the comparative power levels of each of the tests under the various conditions. An apparent trend which appears in the figures involves the RT, Harwell-Serlin and the Adjusted Rank Transform. When there was one nonnull effect (A main effect), the power levels of the unadjusted RT and Harwell-Serlin were similar. However, as soon as a second effect was added, the Blair-Sawilowsky test's power increased, while Harwell-Serlin's power steadily decreased. A review of each graph in each distribution will show the Harwell-Serlin pattern. Harwell-Serlin started approximately equal to the unadjusted RT in each graph, then dropped in each successive graph until the lowest power occurred at the A x B x C interaction. At that point, the graphs show a repetition of the pattern, beginning again with the high point indicated on the A main effect graphs. This is apparent when the sample size was small (n=7), and slightly less apparent when the sample size was large (n=35).

The Gaussian distribution graphs (Figures 1-21) show the power of the ANOVA and the slightly less powerful Blair-Sawilowsky test. It appears from the graphs that very little power was lost when using Blair-Sawilowsky under the Gaussian distribution. The uniform distribution graphs (Figures 22-42) show a similar picture, except the gap between the two tests was slightly wider. The majority of the remaining graphs show the Blair-Sawilowsky test to be more powerful than the ANOVA and remaining tests, and in some instances (Figure 106) considerably more powerful.

Conclusion

The recommendations and conclusions drawn in this section are not to be generalized past the 2 x 2 x 2 layout, or to treatments that impact scale. For example, Blair and Sawilowsky (1990), noted when using their test in a 4 x 3 layout that there were some minor Type I error inflations (i.e., with nominal alpha=.05, Type I error inflated to about .065).

This study indicates that the Analysis of Variance F test shows superior power properties and is robust when used with data from a normal (or uniform) distribution, when compared to Bradley's Collapsed and Reduced Technique, Harwell and Serlin's L test, and Blair and Sawilowsky's Adjusted Rank Transform test. However Micceri (1989) found normal distributions rare when examining real large-sample data sets of test scores.

If the distribution is heavy-tailed or skewed, this study indicates that the Blair-Sawilowsky test statistic is robust and shows superior power properties when compared to ANOVA, Bradley's Collapsed and Reduced Technique, and Harwell and Serlin's L Test. In addition, for the discrete mass at zero data set (real data), Blair-Sawilowsky was much more powerful than ANOVA. When using the other real data set, multi-modal lumpy, Blair-Sawilowsky was slightly more powerful than ANOVA when the effect size was small and slightly less powerful than ANOVA when the effect size was large.

It is recommended that when testing for interactions in a 2 x 2 x 2 layout, Analysis of Variance be used with data known to be symmetric with light tails, such as the normal and uniform distributions, and Blair-Sawilowsky be used with heavy-tailed or skewed data. If the shape of the distribution is unknown, the Blair-Sawilowsky test is recommended because it frequently exhibited considerably more power than the ANOVA, and in the cases when it was less powerful, such as the apparently rare case of normality, it was only slightly less powerful than the ANOVA.

Harwell and Serlin's L test was conservatively and liberally nonrobust when the effect size was large ($\underline{c}$=.8). In addition, as the number of nonnull effects increased, Harwell-Serlin's power decreased. Therefore, it is not recommended that Harwell and Serlin's L test be used in factorial ANOVA layouts.

Bradley's Collapsed and Reduced Technique was usually the least powerful test in this study. It was consistently robust; however, because of its meager power properties, it is not recommended for general purpose use.

Because of the recommendation to use the Blair-Sawilowsky test, Appendix A contains the commands to be used when conducting Blair and Sawilowsky's Adjusted Rank Transform test in SPSS.

Additional Study Areas

There are two remaining tests for interaction identified by Sawilowsky (1990) which were not studied here. Further study may be made of Hettmansperger's (1984) aligned ranks test, and the Shoemaker (1986) test which is based on medians. Additional tests which show promise are the Still and White (1991) test, Berry & Mielke (1983) test, and the Welch (1990) permutation test for interaction (which appeared after the Sawilowsky review). These tests are computationally tedious and require a high level of sophistication in order to perform computation, in part because they are not contained in any computer statistics packages.

Further study may also be made of the promising Blair and Sawilowsky Adjusted Rank Transform test under various other conditions and layouts.

APPENDIX A

The SPSS code for conducting Blair-Sawilowsky's Adjusted Rank Transform test of the interaction is listed below. The sample data set used in this example was the one outlined in Chapter 3, with variables of gender (men and women) and teaching method (A and B). The column and row means are first calculated by using Frequencies, Means, or other SPSS descriptive procedures.

```
data list/gender 1 method 2 Store 3-4.
value labels gender 1 'Male' 2 'Female'/method 1 'A' 2 'B'.
if (gender eq 1 and method eq 1) ajd=score-5.58-9.33.
if (gender eq 1 and method eq 2) ajd=score-5.58-6.08.
if (gender eq 2 and method eq 1) ajd=score-9.33-9.83.
if (gender eq 2 and method eq 2) ajd=score-9.83-6.08.
variable labels ajd 'Aligned Score'.
rank ajd /rank into art.
variable labels art 'Adjusted Rank Transform'.
anova /variables art by gender(1,2) method(1,2).
```

REFERENCES

Akritas, M. G. (1991). Limitations of the rank transform procedure: A study of repeated measures designs, part 1. Journal of the American Statistical Association, 86, 457-460.

Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. Psychological Bulletin, 58, 305-316.

Baker, B. O., Hardych, C. D., & Petrinovich, L. F. (1966). Weak measurement versus strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. Educational and Psychological Measurement, 26, 219-309.

Becker, W. E., & Harnett, D. L (1987). Business and economic statistics with computer application. Reading, MA: Addison-Wesley.

Berenson, M. L., & Levine, D. M. (1992). Basic business statistics: Concepts and applications. (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Berry, K. J., & Mielke, P.W. (1983). Moment approximations as an alternative to the *F* test in analysis of variance. British Journal of Mathematical and Statistical Psychology, 36, 202-206.

Blair, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance." Review of Educational Research, 51(4), 499-507.

Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various nonnormal distributions. Journal of Educational Statistics, 5, 309-335.

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. Psychological Bulletin, 97(1), 119-128.

Blair, R.C., & Sawilowsky, S. S. (1990). A test for interaction based on the rank transform. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Blum, J. R., & Fattu, N. A. (1954). Nonparametric methods. Review of Educational Research, 24, 467-487.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. Psychological Bulletin, 57, 49-64.

Boneau, C. A. (1961). A note on measurement scales and statistical tests. American Psychologist, 16, 160-261.

Box, G. E. P. (1954). Some theories on quadratic forms applied to the study of analysis of variance problems: Effect of unequality of variance in the on-way classification. Annals of Mathematical Statistics, 25, 190-302.

Bradley, J. V. (1968). Distribution-free statistical tests. Englewood Cliffs, NJ: Prentice-Hall.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Bradley, J. V. (1979). A nonparametric test for interactions of any order. Journal of Quality Technology, 11, 177-184.

Bradley, J. V. (1982). The insidious L-shaped distribution. Psychonomic Society, 20(2), 85-88.

Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics, 3, 27-38.

Conover, W. J., & Iman, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs. Communications in Statistics, A5, 1349-1368.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. American Statistician, 35, 124-129.

Conover, W. J., & Iman, R. L. (1982). Analysis of covariance using the rank transformation. Biometrics, 38, 715-724.

Fawcett, R. F., & Salter, K. C. (1984). A monte carlo study of the F test and based on ranks of treatment effects in randomized block designs. Communicating in Statistics, B13, 213-225.

Fisher, R. A. (1935). The design of experiments. Edinburgh: Oliver & Boyd.

Fisher, R. A., & Yates, F. (1938). Statistical tables for biological, agricultural and medical research. Edinburgh: Oliver & Boyd.

Gaito, J. (1960). Scale Classification and Statistics. Psychological Review, 67, 277-278.

Geary, R. C. (1947). Testing for normality. Biometrika, 34, 209-242.

Gibbons, J. D., & Chakraborti, S. (1991). Comparisons of the Mann-Whitney, Student's t, and Alternate t tests for means of normal distributions. Journal of Experimental Education, 258-267.

Glass, G., Peckham, P., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42, 237-288.

Glass, G. V., & Stanley, J. C. (1970). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall.

Gosset, W. S. (Student) (1908a). The probable error of a mean. Biometrika, 6, 1-25.

Gosset, W. S. (Student) (1908b). The probable error of a correlation coefficient. Biometrika, 6, 302-310.

Groggel, D. J. (1987). A Monte Carlo study of rank tests for block designs. Communications in Statistics, 16(3), 601-620.

Harwell M. R. (1988). Choosing between parametric and nonparametric tests. Journal of Counseling and Development, 67, 35-38.

Harwell, M. R. (1991). Completely randomized factorial analysis of variance using ranks. British Journal of Mathematical and Statistical Psychology, 44 383-401.

Harwell, M. R. (1990). A general approach to hypothesis testing for nonparametric tests. Journal of Experimental Education, 58(2), 143-156.

Harwell, M. R. (1991). Summarizing Monte Carlo results in methodological research. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Harwell, M. R. & Serlin, R. C. (1988). An empirical study of a proposed test of nonparametric analysis of covariance. Psychological Bulletin, 104(2), 268-281.

Harwell, M. R. & Serlin, R. C. (1989). A nonparametric test statistic for the general linear model. Journal of Educational Statistics, 14(4), 351-371.

Hettmansperger, T. P. (1984). Statistical inference based on ranks. New York: Wiley.

Higgins, J. J., & Tashtoush, S. (in press). An aligned rank transform test for interaction. Nonlinear World.

Hodges, J. C., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t test. Annals of Mathematical Statistics, 27, 324-335.

Hodges, J. C., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. Annals of Mathematical Statistics, 33, 482-497.

Hollander, M., and Wolfe, D. A. (1973). Nonparametric statistical methods. New York: Wiley.

Hotelling, H. and Pabst, M.R. (1936). Rank correlation and tests of significance involving no assumption of normality. Annals of Mathematical Statistics, 7, 29-43.

Iman, R. L. (1974). A power study of a rank transform for the two-way classification model when interactions may be present. Canadian Journal of Statistics, 2, 227-239.

Iman, R. L., & Conover, W. J. (1976). A comparison of several rank tests for the two-way layout (SAND76-0631). Albuquerque, NM: Sandia Laboratories.

Iman, R. L., Hora, S. C., & Conover, W. J. (1984). Comparison of asymptotically distribution-free procedures for the analysis of complete blocks. Journal of the American Statistical Association, 79, 674-685.

Ingram, J. A., & Monks, J. G. (1992). Statistics for business and economics (2nd ed.). Fort Worth, TX: Dryden.

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In P.R. Krishnaiah (Ed.), Handbook of statistics (Vol. 1, pp. 19-236). The Netherlands: North-Holland.

Keppel, G., & Zedeck, S. (1989). Data analysis for research design. New York: W. H. Freeman and Company.

Kerlinger, F. N. (1973). Foundations of behavioral research. (2nd ed.). New York: Holt, Rinehart and Winston, Inc.

Klugh, H. E. (1986). Statistics: The essentials for research. (3rd ed.). New York: Lawrence Erlbaum.

Mansfield, E. (1986). Basic statistics with applications. New York: W. W. Norton & Company.

Marascuilo, L. A., & McSweeney, M. (1977). Nonparametric and distribution-free methods for the social sciences. New York: Brooks-Cole.

Meddis, R. (1984). Statistics using ranks: A unified approach. Blackwell, England: Oxford.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105(1), 156-166.

Nunnally, J. (1975). Introduction to statistics for psychology and education. New York: McGraw-Hill.

Nunnally, J. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Olejnik, S. F., & Algina, J. (1984). Parametric ANCOVA and the rank transform ANCOVA when the data are conditionally nonnormal and heteroscedastic. Journal of Educational Statistics, 9(2). 129-150.

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. Psychological Bulletin, 83(4), 579-586.

Page, E. B., & Marcotte, D. R. (1966). Nonparametric statistics. Review of Educational Research, 36, 517-528.

Patel, K. M., & Hoel, D. G. (1973). A nonparametric test for interaction in factorial experiments. Journal of the American Statistical Association, 68, 615-620.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. Biometrika, 62(2), 223-241.

Puri, M. L. (1969). Asymptotic efficiency of a class of c-sample tests. Annals of Mathematical Statistics, 35, 102-121.

Puri, M. L., & Sen, P. K. (1971). Nonparametric methods in multivariate analysis. New York: Wiley.

Rasmussen, J. L. (1985). The power of student's t and Wilcoxon W statistics. Evaluation Review, 9(4), 505-510.

Runyon, R. P., & Haber, A. (1991). Fundamentals of behavioral statistics (7th ed.). New York: McGraw-Hill.

Sawilowsky, S. S. (1985). Robust and power analysis of the 2 x 2 x 2 ANOVA, rank transformation, random normal scores, and expected normal scores transformation tests. (Unpublished doctoral dissertation, University of South Florida, Tampa, FL).

Sawilowsky, S. S. (1989, April). Rank transform: The bridge is falling down. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. Review of Educational Research, 60(1), 91-126.

Sawilowsky, S. S. (1993). Comments on using alternatives to normal theory statistics in social and behavioral science. 34(4), 432-439.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. Psychological Bulletin, 111(2), 352-360.

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA. Journal of Educational Statistics, 14(3), 255-267.

Scheffe', H. (1959). The analysis of variance. New York: Wiley.

Senders, V. L. (1958). Measurement and Statistics. New York: Oxford University Press.

Shaffer, J. P. (1977). Reorganization of variables in analysis of variance and multidimensional contingency tables. Psychological Bulletin, 2, 220-228.

Shoemaker, L. H. (1986). A nonparametric method for analysis of variance. Communications in Statistics, 15(3). 609-632.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103, 677-680.

Still, A. W., & White, A. P. (1981). The approximate randomization test as an alternative to the F test in analysis of variance. British Journal of Mathematical and Statistical Psychology, 34, 243-252.

Tate, M. W., & Clelland, R. C. (1957). Nonparametric and shortcut statistics. Danville, IL: Interstate.

Thompson, G. L. (1991). A unified approach to rank tests for multivariate and repeated measures designs. Journal of the American Statistical Association, 86(414), 410-419.

Toothaker, L. E., Chang, H., (1980). On "the analysis of ranked data derived from completely randomized factorial designs." Journal of Educational Statistics, 5(2), 169-176.

Welch, W. J. (1990). Construction of permutation tests. Journal of the American Statistical Association, 85(411), 693-698.

Wilson, K. V. (1956). A distribution-free test of analysis of variance hypotheses. Psychological Bulletin, 53(1), 96-101.

Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

ABSTRACT

THE COMPARATIVE POWER OF SEVERAL NONPARAMETRIC ALTERNATIVES TO THE ANALYSIS OF VARIANCE TEST FOR INTERACTION IN A 2 x 2 x 2 LAYOUT

by

DEBORAH LYNN KELLEY
MAY 1994

Advisor: Shlomo Sawilowsky
Major: Evaluation and Research
Degree: Doctor of Philosophy

Historically, there has been a lack of nonparametric tests available for use in detecting interactions in Analysis of Variance. Over the past 25 years, new nonparametric tests for interactions have been developed. The purpose of this study was to determine the comparative robustness with respect to departures from normality (i.e., Type I error) and power properties of four tests of interaction in the 2 x 2 x 2 ANOVA layout. The four tests were: a) Analysis of Variance, a parametric test; b) Bradley's Collapsed and Reduced Technique, a distribution-free test; c) Harwell and Serlin's L test, which is based on the Puri and Sen (1985) test; and d) Blair and Sawilowsky's Adjusted Rank Transform test, which is similar to Fawcett and Salter's (1984) aligned rank test, and is a conditionally distribution-free test. The study used Monte Carlo techniques to compare the four tests under four theoretical distributions and two real (psychological and educational) data sets from Micceri (1989). Sample sizes of 7, 21, and 35 and effect sizes of .2 - 1.2(.2) were used.

The results indicated that the Analysis of Variance F test shows superior power

properties and is robust with data from symmetric and light tailed distributions, when compared to Bradley's Collapsed and Reduced Technique or Harwell and Serlin's L test. Blair and Sawilowsky's Adjusted Rank Transform test, however, was only slightly less powerful than ANOVA under these conditions.

If the distribution is heavy-tailed or skewed, the results indicated that the Blair-Sawilowsky statistic is robust and shows superior power properties when compared to ANOVA, Bradley's Collapsed and Reduced Technique, and Harwell and Serlin's L Test. In addition, when using a real psychometric data set, discrete mass at zero, Blair-Sawilowsky was much more powerful than ANOVA. When using a real achievement data set, multi-modal lumpy, Blair-Sawilowsky was slightly more powerful when the effect size was small and slightly less powerful than ANOVA when the effect size was large.

Harwell and Serlin's L test was conservatively and liberally nonrobust when the effect size was large. When the number of non-null effects increased from one to seven, Harwell-Serlin's power decreased. Bradley's Collapsed and Reduced Technique was usually the least powerful test in this study. It was consistently robust. However, it demonstrated meager power properties. For these reasons, neither the Harwell-Serlin or Bradley test is recommended.

AUTOBIOGRAPHICAL STATEMENT

D. LYNN KELLEY

WORK EXPERIENCE

1/94 - Present	**MADONNA UNIVERSITY, 36600 Schoolcraft, Livonia, MI 48150**
	Associate Professor and Chairperson, Quantitative Methods and Computer Systems Department, School of Business
11/92 - 1/94	**WAYNE STATE UNIVERSITY, 326 Justice Building, Detroit, MI 48202**
	Assistant Director/Research Associate, Developmental Disabilities Institute
1984 - 1992	**DOCTORS HOSPITAL, 2730 E. Jefferson, Detroit, MI 48207**
2/90-2/92	Executive Vice President and Chief Operating Officer
4/89-2/90	Vice President, Operations
5/87-4/89	Leasing & Operations Manager/Administrative Services Manager
6/85-5/87	Project Information Coordinator
2/84-5/87	Administrative Assistant to the President

EDUCATION

ABD	Wayne State University Major: Evaluation and Research (Statistics)	GPA: 4.0	1991-1994
MBA	Michigan State University Executive MBA Program		1990 Graduated
BBA	University of Detroit Major: Business Management	Summa Cum Laude	1988 Graduated

Three week international business study course held in Japan and Korea: 1989