DOCUMENT RESUME

ED 365 721 TM 020 943

AUTHOR Daniel, Thomas Dyson
TITLE A Statistical Power Analysis of the Quantitative Techniques Used in the "Journal of Research in Music Education," 1987 through 1991.
PUB DATE Nov 93
NOTE 26p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (New Orleans, LA, November 10-12, 1993).
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Effect Size; *Estimation (Mathematics); *Music Education; Nonparametric Statistics; *Research Design; Research Methodology; *Sample Size; Sampling
IDENTIFIERS A Priori Tests; *Journal of Research in Music Education; *Power (Statistics)

ABSTRACT

Statistical power in music education was examined by taking an in-depth look at quantitative articles published in the "Journal of Research in Music Education" between 1987 and 1991, inclusive. Of the 109 articles of the period, 78 were quantitative, with both parametric and nonparametric procedures considered. Sample sizes were those reported by the authors. Effect sizes were estimated according to the guidelines developed by J. Cohen (1988), and his power analysis tables were used. The overall median power for the articles was 0.13 for detecting small effects, 0.64 for detecting medium effects, and 0.97 for detecting large effects. Implications of these findings, limitations of the study, and suggestions for future music-education research are discussed. In general, more attention should be placed on a priori power analyses of research designs. Adequate sample sizes should be chosen and a greater understanding and application of the concept of effect size is needed in music education research. Six tables are included. (Contains 18 references.) (SLD)

A STATISTICAL POWER ANALYSIS OF THE QUANTITATIVE TECHNIQUES USED IN THE *JOURNAL OF RESEARCH IN MUSIC EDUCATION*, 1987 THROUGH 1991

Thomas Dyson Daniel

Auburn University

Dissertation presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, Louisiana, November, 1993.

INTRODUCTION

The use of statistics in music education research seems to have grown substantially with the increasing availability of desktop computers. Other branches of the behavioral sciences have experienced this same trend and have produced methodological studies of their respective fields. These studies and assessments in other fields are aimed at improving the procedures and statistical methods and often include a power analysis of quantitative research recently published.

In music education research, a power analysis study has never been published. The overwhelming majority of methods, procedures, and movements attempted in music education today are supported by research published in reputable and learned journals. Additionally, the majority of the published research is quantitative in nature, employing a vast array of statistical procedures unheard of in music education not long ago. Yet, as the availability of sophisticated techniques increases are music educators paying attention to the problem of statistical power?

Neyman and Pearson (1928) created the concept of power in two articles published in *Biometrika*, where they introduced the idea of designing research with more than a simple null hypothesis to test. Neyman and Pearson outlined the existence of two hypotheses to choose between, the null and the alternative, and also proposed a risk level associated with the alternative hypothesis, called beta, which would serve the same function as the alpha risk served with the null hypothesis. This construction led directly to the concept of Type II error. Neyman and Pearson recognized that if rejecting a true null hypothesis is an error of the first kind controlled by alpha, then failing to reject the null hypothesis when the alternative hypothesis is true must be an error of the second kind controlled by beta. In addition, Neyman and Pearson argued that if a researcher set the effect size, alpha risk, and sample size, then he could determine beta and the consequent probability of rejecting the null hypothesis, which is called statistical power.

Cohen (1962) successfully rekindled the idea of power analysis in his classic study where he calculated a statistical power for three effect sizes for each of the 70 quantitative articles published in the 1960 volume of *Journal of Abnormal and Social Psychology* (now *The Journal of Abnormal Psychology*). Cohen distinguished between major and minor research hypotheses within each study and suggested three levels of effect size (small, medium, and large) to be separately defined by the nature of the specific statistical test that was used. Combining each calculated statistical power at the nondirectional .05 level, regardless of the alpha level actually reported, Cohen derived a median power for each of the three effect sizes. For medium effects Cohen reported a median power of .46, which

would indicate that these studies had slightly less than a 50-50 chance of detecting medium effects.

In 1988, Cohen issued the second edition of his *Statistical Power Analysis for the Behavioral Sciences* with the most prominent change being the addition of a chapter for power analysis of set correlation and multivariate methods. Also, some mistakes in the earlier editions were corrected, some power and sample size tables for multiple regression and correlation were updated, and a new chapter was added dealing with practical issues in current power analysis. Borenstein and Cohen (1988) published *Statistical Power Analysis: A Computer Program*, which apparently serves in a complementary role to Cohen's (1988) power handbook. This program follows the table of contents of Cohen's power handbook almost exactly as the screen menu for each module lists procedures for *t*-tests, correlations, differences between proportions, one-way analysis of variance, factorial analysis of variance, and multiple regression/correlation. Each module also offers several internal options, including power or sample size computations, effect size computations, tables and graphs, and Monte Carlo simulations. However, omitted from this computer program is Cohen's new chapter 10 from his power handbook relating to the power analysis of set correlation and multivariate methods.

Sedlmeier and Gigerenzer (1989) recreated Cohen's classic 1962 power analysis study in which they calculated the median power of the 1984 volume of the *Journal of Abnormal Psychology* for the purpose of comparison as to whether or not other similar studies of statistical power had any effect on the statistical power of behavioral science studies. Their calculations show that after 24 years the median power declined rather than increased. The authors also pointed out the importance of having faith in the truth of the basic assumptions of the statistical model, whether or not they are actually stated and noted the purely intuitive judgment used by Cohen in estimating the values of his three effect sizes. They further noted that these effect size values seem to have withstood the test of time with only minor alterations.

Rossi (1990) also duplicated Cohen's 1962 power analysis survey by analyzing the 1982 volume of the *Journal of Abnormal Psychology*, in addition to the *Journal of Consulting and Clinical Psychology* and the *Journal of Personality and Social Psychology*. The most intriguing aspect of Rossi's study was his Table 1 on page 648 in which he compared the final results of 25 separate power analysis surveys covering 40,000 statistical tests and over 1,500 journal articles. In combining all of the aforementioned power analysis surveys, Rossi reported that, "The average statistical power for all 25 power surveys (including Cohen's) was .26 for small effects, .64 for medium effects, and .85 for large effects," (Rossi, 1990, p. 647).

Cohen (1990) showed the real danger inherent in a study with low power when a failure to reject the null hypothesis is mistakenly interpreted as a verification of the null hypothesis. For example, a highly respected researcher in any field might use his influence and the failure to reject the null hypothesis as a tool to destroy an important theory of learning. Cohen (1990) related a personal experience when a colleague conducted a study with a sample size of 20 cases per group, found a nonsignificant result, and "proceeded to demolish an important branch of psychoanalytic theory" (Cohen, 1990, p. 1304). Cohen later calculated his friend's power to have been .33 for medium effects.

Daniel (1992) evaluated all statistical procedures used in the *Journal of Research in Music Education* from 1987 through 1991. This study revealed no mention of statistical power or effect size in any article published in *JRME* during this time. Also, a tally of all statistical techniques, both parametric and non-parametric included, and sample sizes was completed, and the statistical techniques were divided into categories of Basic, Intermediate, Advanced, and Other according to the degree of difficulty of each statistical technique. Results for the individual articles were grouped into issue totals, volume totals, and overall item totals.

Of the 109 articles published in the *JRME* from 1987 through 1991, 78 were quantitative, or 72%, and 31 were non-quantitative, or 28% (Daniel, 1992). Perhaps the most surprising aspect of the study concerned the two items of Power and Effect Size. Of the 78 quantitative articles, none reported any type of power analysis or any mention of a power estimate or estimate of effect size. In fact, the words "power" and "effect size" were never encountered anywhere in the context of Type II error. The pilot study also calculated a median sample size of N=99 per study. In cases where separate sample sizes were used within a single article each different sample size was treated as a separate unit. The smallest sample size encountered was N=12 in 1991 and the largest was N=1,843 in 1990. The results of the sample size calculations are reported below in Table 1.

Insert Table 1 about here

A power analysis of music education research would create awareness of the importance of statistical power in music education, would provide the opportunity for assessments of the attention paid to Type II error, effect size, and statistical power in music education, could lead to a discussion of editorial practices for the *JRME*, and could open the door for many possible similar studies of other music education publications.

Consider the structure of a music education manuscript accepted for publication by a research journal. If this music education research article contained mention of statistical

power, then the reader could safely assume that the author had experienced at least a brief consideration of Type II error and effect size. However, if there was no mention of statistical power, then the reader must conclude one of two explanations for the omission: the author ignored Type II error and effect size, or the editor of the journal elected to remove this section of the article due to editorial policy. Either of these possibilities suggests an ignorance of the importance of statistical power in music education research.

On a larger scale, consider not just one article but 100 hypothetical articles of music education research. Finding mention of statistical power in every single quantitative article would be both astounding and wonderful. With an incredible display of statistical awareness, perhaps a majority of these 100 articles might contain mention of statistical power. However, consider the possibility that the majority of these 100 articles display an ignorance of statistical power. If these are indeed the current conditions of research in music education, then this study should be welcome in the research community to begin the process of the assessment of research methodologies. This study and its influence is long overdue if one considers the unfortunate possibility that none of these 100 hypothetical articles contained any mention of statistical power or effect size.

Purpose of the Investigation

The purpose of this study was to examine statistical power in music education research by taking an in-depth look at quantitative articles published in the *Journal of Research in Music Education* (*JRME*), a publication of the Music Educators National Conference. This study calculated and examined the calculated post-hoc statistical power of every principle technique used in each quantitative article published in the *JRME* between 1987 and 1991, inclusive.

This study is important for several reasons. First, this study should serve as a reminder that in order for music education research to be taken seriously by consumers and other behavioral scientists, it must display all the requirements of standard scientific method. By means of comparison, researchers can instantly know how music education fares along with other areas of the behavioral sciences, and by turning more attention to studies of statistical power researchers in the field may routinely begin to consider statistical power during the planning stages.

Second, this study could serve as a catalyst to encourage music educators to refine and improve their statistical design norms and habits. If statistical power is discovered to be too low, guidelines could be suggested to create research designs which have a better probability of detecting significant differences that might exist. Consequently, if statistical

power is discovered to be too high, music education researchers could learn how to trim the research design to be more cost-efficient while maintaining adequate power.

Third, a search of current behavioral science literature has revealed a lack of studies of statistical power in music education. Therefore, the results of this study should add to the understanding of the design of music education research and lead to improvements in future research, whether the existing designs are adequate or not. Music educators are, unfortunately, handicapped in their training as empiricists. According to Rainbow and Froehlich (1987), music educators are trained first and foremost as musicians, and the traditional music lessons and pedagogical methods are dominated by an authoritative expert or master. Musicians are instructed at an early age to accept as fact the word of the master. "Because of the conditioned admiration of and respect for the expert, musicians sometimes find it difficult to trust their own judgements and insights into things. They hesitate to break away from the mold of expert experiences and to raise questions about the veracity of an authority statement or one derived from common sense," (Rainbow and Froehlich, 1987, p. 26). Although this method of instruction has worked well for musicians, it does not help create free-thinking, doubting, and inquisitive researchers and empiricists.

METHOD

The *Journal of Research in Music Education*, created in 1953 by the Music Educators National Conference, continues to be the main avenue of support for research in music education and would be considered the primary research publication in the field of music education (Mark, 1986). Over the past five years, the journal has been a publication of largely quantitative content and should provide many quantitative articles from which to choose. In fact, the editor of *JRME*, Rudolf E. Radocy, defended the quantitative nature of the publication against apparent charges of bias toward quantitative research by saying, "The Editorial Board...has no systematic bias against any type of research. Admittedly, many articles are quantitative. The Editorial Board welcomes all types of research...but we cannot publish articles that are not submitted" (Radocy, 1988, p. 204). As Radocy explained, if the majority of manuscripts submitted to the *JRME* are quantitative then the nature of the journal will be mostly quantitative.

Procedure

The design for this study was first attempted by Jacob Cohen in 1962 and has been duplicated by many other researchers in many other fields. The data used in this study were collected from the principal statistical procedures used in all quantitative articles published in the *Journal of Research in Music Education* during the years 1987 through

1991. There was a total of 109 articles published in these five volumes of *JRME*, of which 78 were quantitative and 31 were non-quantitative. In addition, both parametric and non-parametric procedures were considered for analysis.

The main instrument used in this study was J. Cohen's power analysis tables found in the 1988 edition of his text, *Statistical Power Analysis for the Behavioral Sciences* (Cohen, 1988). Also, some calculations were double-checked using Borenstein and Cohen's 1988 power analysis computer program entitled *Statistical Power Analysis: A Computer Program* (Borenstein and Cohen, 1988). The computer program provides power tables and sample size tables for the following procedures; *t*-tests, correlations, differences between proportions, chi square, one-way and factorial analysis of variance, one-way and factorial analysis of covariance, and multiple regression/correlation (Borenstein and Cohen, 1988).

The sample sizes used in the analysis were those reported by the authors of the articles in question. The effect sizes were estimated according to Cohen's guidelines (1988) as listed below in Table 2. Usually, the author's own stated effect size should be used in post-hoc power analysis, but no articles in the *Journal of Research of Music Education* from 1987 through 1991 stated an estimated effect size. For this reason, all three estimated effect sizes were used and three separate statistical powers were calculated for each procedure analyzed. The alpha level used was always the nondirectional .05 level, as recommended by Cohen (1962), regardless of the actual alpha level stated by the author.

Insert Table 2 about here

For the few instances where non-parametric procedures appeared, each procedure was equated with its analogous parametric test (Cohen, 1962). For example, the Mann-Whitney U test and the Wilcoxin matched-pairs signed-ranks test were calculated as if they were the *t*-test for means (Cohen, 1962). Additionally, the Kruskal-Wallis one-way analysis of variance and Friedman's two-way analysis of variance were treated as if they were a simple *F*-test (Cohen, 1962). Also, this study was one of the first to use Cohen's guidelines for multivariate power analyses since earlier studies were forced to ignore multivariate power analyses because of a lack of computational procedures. The newest edition of Cohen's power analysis textbook (Cohen, 1988) includes a new chapter on multivariate techniques making multivariate power analysis calculations more accessible.

After determining the parametric and non-parametric procedures to be examined further, the statistical techniques were grouped according to their relationship with the major research hypotheses in the study. As suggested by Sedlmeier & Gigerenzer (1989),

only the procedures relating to the major research hypotheses were examined. These techniques were grouped into units separated by sample size (each different sample would be part of a new unit). Using the authors' given sample sizes, nondirectional .05 alpha, and estimated effect sizes, a post-hoc statistical power was calculated for each technique, according to Cohen's procedures and power tables (Cohen, 1988). This information was then combined and examined in a variety of ways.

First, an overall median power was calculated for each effect size to represent an overview of the study. Second, additional information was provided by examining the median power of each individual volume of the *JRME*. This information was compared to the overall median power to determine the existence or nature of any possible trends over the five-year period.

Third, the median power for each statistical technique used in the articles was calculated for each effect size. For example, the median power for all multiple regression procedures was calculated for each effect size, the median power for all canonical correlation procedures was calculated for each effect size, the median power for all chi square procedures was calculated for each effect size, etc. This information provided a basis for comparison of the statistical powers of the different categories of procedures.

Finally, these procedures were identified as either Basic, Intermediate, Advanced, or Other procedures, and a median power for each group was calculated for each effect size. The groupings proceeded as follows: Basic techniques included one-way analysis of variance (ANOVA), chi square, Pearson correlation, *z*-test, and *t*-test; Intermediate procedures included one-way analysis of covariance (ANCOVA), factorial ANOVA/ANCOVA, and multiple regression. Advanced procedures included one-way and factorial multivariate analysis of variance (MANOVA), one-way and factorial analysis of covariance (MANCOVA), discriminant analysis, and canonical correlation. Other procedures included any other non-parametric procedures, such as the Spearman rank, Mann-Whitney U test, the Wilcoxin matched-pairs signed-ranks test, the Kruskal-Wallis one-way analysis of variance, Kendall's coefficient of concordance W, and Friedman's two-way analysis of variance.

Several delimitations were necessary in order to narrow the focus of this study. First was the time frame of articles to be considered. Other similar studies have used time frames ranging from one to four years. The years 1987 through 1991 were chosen because they represented the five most recently published and completed volumes of the *JRME* at the time of this study. Each of the other similar studies used some rationale for their selection of a time frame, and this study was designed to examine at least 100 articles. Since the volumes of 1987 through 1991 contained 109 total articles, this was considered a

suitable selection. A second delimitation of this study was the use of quantitative articles for power analysis. The need for this delimitation is required because non-quantitative articles, such as historical, qualitative, methodological, or philosophical, cannot be analyzed for statistical power because they involved no testing of the null hypothesis.

Initially, another delimitation of this study was the use of strictly parametric procedures. However, closer investigation revealed that non-parametric procedures could be treated in the same way as similar parametric procedures (Cohen, 1962). For example, the Mann-Whitney U test and the Wilcoxin matched-pairs signed-ranks test were calculated as if they were the *t*-test for means (Cohen, 1962). Additionally, the Kruskal-Wallis one-way analysis of variance and Friedman's two-way analysis of variance were treated as if they were a simple *F*-test (Cohen, 1962). Also, since Kendall's coefficient of concordance W was reported as a chi square value, then the corresponding statistical power was calculated identically to a chi-square procedure. Therefore, both parametric and non-parametric procedures were considered for power analysis.

A third delimitation of this study was the use of Cohen's suggested effect size values of small, medium, and large. Ideally, an author should state the estimated effect size of his own study. However, since a recent survey found no mention of effect size in the articles in question (Daniel, 1992), effect sizes were estimated according to Cohen's guidelines (Cohen, 1988). A fourth delimitation of this study was the use of the nondirectional .05 level, regardless of the alpha level actually reported in the study. This procedure, taken directly from Cohen (1962) allowed a direct comparison among different studies.

A fifth delimitation of this study was the use of the median as the reported measure of central tendency. The median represents the value at which the distribution is divided into the upper and lower 50 percent of the scores (Shavelson, 1988). This particular measure was chosen because of the similar design of other power analysis studies which allowed a direct comparison among all similar studies. A sixth delimitation of this study was the use of statistical procedures only associated with the major research hypotheses. Many articles frequently used statistical procedures in pilot studies or reliability computations, and these results were reported in the published article. However, according to Sedlmeier and Gigerenzer (1989), the final power calculations were not equally comparable unless the procedures used in the major research hypotheses were the only ones considered.

RESULTS

The overall median statistical power for the *JRME*, 1987 through 1991, was calculated to be .13 for small effects, .64 for medium effects, and .97 for large effects. For each volume, the median statistical powers were calculated to be as follows: 1987, .18 for small effects, .86 for medium effects, and 1.00 for large effects; 1988, .12 for small effects, .66 for medium effects, and .99 for large effects; 1989, .11 for small effects, .50 for medium effects, and .905 for large effects; 1990, .125 for small effects, .65 for medium effects, and .98 for large effects; and 1991, .13 for small effects, .66 for medium effects, and .97 for large effects. These results are noted below in Tables 3A and 3B.

Insert Tables 3A and 3B about here

Since ANOVA was the dominant statistical procedure in the *JRME* from 1987 through 1991, a closer look at that procedure is warranted followed by examples of the power calculations. The overall median statistical power for ANOVA procedures in the *JRME*, 1987 through 1991, was calculated to be .11 for small effects, .51 for medium effects, and .91 for large effects. For each type of ANOVA, the median statistical powers were calculated as follows: one-way ANOVA, .125 for small effects, .60 for medium effects, and .96 for large effects; factorial ANOVA, .11 for small effects, .50 for medium effects, and .905 for large effects; and ANCOVA, .11 for small effects, .56 for medium effects, and .96 for large effects. These results are noted below in Tables 4A and 4B.

Insert Tables 4A and 4B about here

The following example of a one-way ANOVA power calculation is provided from volume 36, number 4, pages 205-219 of the *JRME* (Kendall, 1988). In this procedure the alpha level is .05, the total number of subjects is 76, and 4 groups are used with n=19 subjects per group. The degrees of freedom of the numerator of the F ratio in ANOVA is calculated as the number of groups minus one, or u=3. The effect sizes of ANOVA procedures are listed in Section 8.2.3 (Cohen, 1988) as f=.10 for small effects, f=.25 for medium effects, and f=.40 for large effects. Looking at Table 8.3.14 (Cohen, 1988), the statistical power for n=19 is .09, .41, and .83 for small, medium, and large effects.

The following example of a factorial ANOVA power calculation is provided from volume 37, number 1, pages 5-20 of the *JRME* (Kratus, 1989). In this procedure, the alpha level is .05, the total number of subjects is 60, and the design is a 3 X 2 factorial design. The effect sizes of ANOVA procedures are listed in Section 8.2.3 (Cohen, 1988) as f=.10 for small effects, f=.25 for medium effects, and f=.40 for large effects. The

degrees of freedom of the numerator of the F ratio in ANOVA is calculated as the number of levels minus one, or u=2 for Factor A, u=1 for Factor B, and u=2 for the interaction term A X B. Using Formula 8.3.4 (Cohen, 1988), the n' for each factor is calculated as n'=19 for Factor A, n'=28 for Factor B, and n'=19 for A X B. Looking at Table 8.3.13 (Cohen, 1988), the statistical power for Factor A at n'=19 is .09, .36, and .76 for small, medium, and large effects. Looking at Table 8.3.12 (Cohen, 1988), the statistical power for Factor B at n'=28 is .11, .46, and .84 for small, medium, and large effects. Looking at Table 8.3.13 (Cohen, 1988), the statistical power for Factor A X B at n'=19 is .09, .36, and .76 for small, medium, and large effects.

Since power calculations for multivariate procedures are relatively new, a closer look at those procedures is warranted followed by examples of the power calculations. The overall median statistical power for all multivariate procedures in the *JRME*, 1987 through 1991, was calculated to be .15 for small effects, .87 for medium effects, and 1.00 for large effects. For each type of multivariate procedure the median statistical powers were calculated as follows: MANOVA was .11 for small effects, .72 for medium effects, and 1.00 for large effects; discriminant analysis was .17 for small effects, .92 for medium effects, and 1.00 for large effects; MANCOVA was .185 for small effects, .815 for medium effects, and 1.00 for large effects; and canonical correlation was .18 for small effects, .95 for medium effects, and 1.00 for large effects. These results are noted below in Tables 5A and 5B.

Insert Tables 5A and 5B about here

The following example of a MANOVA power calculation is provided from volume 37, number 4, pages 258-271 of the *JRME* (Schmidt, 1988). In this procedure, the alpha level is .05 and the total number of subjects is N=43. The design contained 8 dependent variables and 4 independent variables for a total of 12 variables. The value of k for the dependent variables is 8 and the value of k for the independent variables is (4-1), or 3. Table 10.2.1 (Cohen, 1988) gives s=2.43, and the effect sizes of MANOVA procedures are listed in Section 10.2.2 (Cohen, 1988) as f^2=.02 for small effects, f^2=.15 for medium effects, and f^2=.35 for large effects. From equation 10.1.6 (Cohen, 1988) the numerator degrees of freedom is the two k values multiplied together, or u=(8)(3) or u=24. From equation 10.1.8 (Cohen, 1988) the value of m=43 - (8 + 4 + 3) / 2=35.5, and from equation 10.1.7 (Cohen, 1988) the denominator degrees of freedom is v=(35.5)(2.43) + 1 - (32/2)=71.265. The values of lambda from Equation 10.3.1 (Cohen, 1988) are 1.92, 14.44, and 33.69 for small, medium, and large effects. Looking at Table 9.3.2 (Cohen,

1988), the statistical power for u=24 and v=60 is 0 for lambda=0 and .08 for lambda=2, .40 for lambda=14 and .46 for lambda=16, and .84 for lambda=32 and .89 for lambda=36. Interpolating between these figures, the statistical power for u=24 and v=60 is .07, .41, and .86 for lambdas of 1.92, 14.44, and 33.69. Looking again at table 9.3.2 (Cohen, 1988) the statistical power for u=24 and v=120 is 0 for lambda=0 and .08 for lambda=2, .46 for lambda=14 and .53 for lambda=16, and .90 for lambda=32 and .94 for lambda=36. Interpolating between these figures, the statistical power for u=24 and v=120 is .07, .47, and .91 for lambdas of 1.92, 14.44, and 33.69. Interpolating for v=71.265, the statistical power is .07, .42, and .87 for small, medium, and large effects.

As noted earlier, each procedure was classified according to its degree of difficulty. Basic techniques include one-way analysis of variance (ANOVA), chi square, Pearson correlation, and *t*-test; Intermediate procedures include one-way analysis of covariance (ANCOVA), factorial ANOVA/ANCOVA, and multiple regression. Advanced procedures include one-way and factorial multivariate analysis of variance (MANOVA), one-way and factorial analysis of covariance (MANCOVA), discriminant analysis, and canonical correlation. Other procedures include any other non-parametric procedures, such as the Mann-Whitney U test, the Wilcoxon matched-pairs signed-ranks test, the Kruskal-Wallis one-way analysis of variance, Kendall's coefficient of concordance W, and Friedman's two-way analysis of variance.

The statistical power for Basic procedures was calculated to be .13 for small effects, .68 for medium effects, and .99 for large effects. The statistical power for Intermediate procedures was calculated to be .11 for small effects, .52 for medium effects, and .91 for large effects. The statistical power for Advanced procedures was calculated to be .15 for small effects, .87 for medium effects, and 1.00 for large effects. The statistical power for Other procedures was calculated to be .15 for small effects, .67 for medium effects, and .98 for large effects. These results are noted below in Table 6A.

Insert Table 6A about here

Comparing these procedural results across a medium effects size would yield .68, .52, .87, and .67 for each consecutive degree of procedure listed above. There were 255 total power calculations for each effect size. For each degree of procedure, the total power calculations for each effect size were 107 Basic procedures, 109 Intermediate procedures, 18 Advanced procedures, and 21 Other procedures. The results of a final comparison of

medium powers across all statistical procedures can be seen in Table 6B.

Insert Table 6B about here

CONCLUSIONS

As shown in Table 3A, the overall median power for the quantitative articles published in the *JRME* from 1987 through 1991 is .13 for detecting small effects, .64 for detecting medium effects, and .97 for detecting large effects. Interestingly, the results of this study are very close to Rossi's average statistical power for 25 separate power analysis surveys in which he reported .26 for small effects, .64 for medium effects, and .85 for large effects (Rossi, 1990).

An examination of the statistical power of the Degree of Difficulty groupings found in Table 6A shows the comparison among the different groupings. As expected, the Advanced procedures displayed the highest statistical power (.87) of all groups. Surprisingly, the Intermediate procedures (.52) were lower in power than the Basic procedures (.63). This can probably be explained by the relatively high power of the Basic chi square and Pearson correlation procedures and the relatively low power of the Intermediate factorial ANOVA and ANCOVA procedures.

With an overall median power of .64, authors contributing to the *JRME* from 1987 through 1991 had less than a two-thirds chance of rejecting the null hypotheses from their major research questions (unless they were seeking large effects). The statistical power figures derived from this study were in the upper range of the .40 to .70 framework hypothesized at the beginning of this study. Most of the other similar power analysis studies reported a median statistical power within this range. Therefore, since this study uncovered results that were similar to other studies, perhaps the reasons for the low statistical power are also similar.

To begin with, the original parameters of power analysis (alpha level, effect size, and sample size) should be examined for their possible contributions to the inadequate statistical power uncovered in this study. According to the methods of this power analysis study, all alpha levels were set at the .05 level in order to make the final results more comparable across the board. In only one article was it necessary to lower the original alpha level to .05 for the purpose of this study. In all other cases where an adjustment in alpha level occurred the alpha had to be raised to the .05 level. According to the parameters of power analysis, raising alpha has the effect of raising power, and in cases where alpha was raised this must have had a beneficial effect on the actual statistical power. If all alpha

levels were set at the .05 level for this study (regardless of their true level), alpha can be eliminated as a factor in the inadequate power uncovered.

As already noted, none of the 78 quantitative articles contained any mention of the crucial issue of effect size, the heart of power analysis. Effect size is not as restricted as alpha level, which is usually constrained to the .01, .05, or .10 level, or desired power, which is usually set at the .80 level. Effect size can be set at any level the researcher wishes, as long as the researcher has justification. As Cohen points out, if a researcher wants to know what his chances are of finding something, "the researcher needs to have some idea of how big 'it' is" (Cohen, 1988, p. 532). Type II error cannot possibly be committed if the null hypothesis is rejected. Therefore, low power can never cast doubt on studies producing statistically significant results, and significant results discovered in a study with low power could be the result of three things: an inaccurate estimation of effect size; luck; or both.

If the influence of effect size on statistical power is indeed so crucial, what prevents researchers from simply reporting large effect sizes in every study simply to gain high power? Creating a deceptively high effect size in order to falsify statistical power should be considered similar to a falsification of data and a serious ethics violation.

The overall median sample size for the *JRME* from 1987 through 1991 was calculated to be N=99. Looking at Table 1, the optimum sample size was N=145 in order to meet the conditions of alpha=.05, medium effect size, and a power of .80. This optimum sample size was calculated by using Cohen's sample size tables (Cohen, 1988) for each statistical test. Table 1 can also offer another explanation for the sudden drop in statistical power for the 1989 volume where the true sample sizes used in the statistical tests for that year were less than half (47%) of what they should have been in order to achieve a power of .80 for medium effects.

Recommendations for Future Research

What effect does low power have on any future research in the field of music education? First, low power inhibits the chances of a significant result. Perhaps there is an editorial bias in the *JRME* towards "successful" research. "Successful" research would be any research reporting a significant difference or significant correlation (rejecting the null) and "unsuccessful" research would be any research not reporting a significant difference or significant correlation (failing to reject the null). Therefore, conducting a study with high statistical power would seem to automatically increase the chances of publication in *JRME* based on an increase in the chances of "success". If the initial research design is rigorously planned, there are fewer chances of inconsistent results.

Without exception, each quantitative article examined for this study contained a significant difference or a significant correlation in at least one major research hypothesis. This problem seems to stem from an editorial push toward publishing only significant results. Although music educators can be commended for formulating hypotheses that were validated by their research, what happened to all the well-designed research that uncovered no significant differences or correlations? Three possibilities are suggested to explain this occurrence: (a) Every research design formulated in music education miraculously contained at least one major research hypothesis with a significant result; (b) Research designs in music education that failed to uncover any significant results were rejected for publication on those grounds; (c) Research designs in music education that failed to uncover any significant results were never submitted for publication for fear of rejection. Although the miraculous effects of music on some people cannot be doubted, perhaps the actual research designs in music education were somewhat less than perfect, and the first possibility above can be dismissed.

If the first possibility can be dismissed, then the blame for significance bias in music education research can be placed at the doorstep of editorial policy. If editors are reluctant to publish studies that fail to uncover significance, as outlined in the second possibility, then they are indirectly creating the atmosphere outlined in the third possibility that discourages article submission on the grounds of the absence of significant results. The researchers probably will not submit anything that they feel would not be published, yet the editors cannot possibly publish anything that is not submitted. If this is truly the situation at hand, then an unfortunate downward spiral has already begun, as each condition feeds the other. As Jacob Cohen pointed out, "It seems obvious that investigators are less likely to submit for publication unsuccessful than successful research, to say nothing of a similar editorial bias in accepting research for publication" (Cohen, 1962, p. 151).

Therefore, in order to have the best chances of avoiding the whole problem and eliminating the downward spiral, researchers should simply pay more attention to statistical power in the initial research design. Ideally, the editorial board of the *JRME* could be encouraged to set new guidelines and standards based partially on attention to adequate statistical power and effect size for the acceptance of quantitative manuscripts to be published. "Any investigation in music education that touches on the concerns of another field of study must be conducted in a way that meets professional criteria acceptable to both music education and the external area," (Rainbow & Froehlich, 1987, p. 30).

Another effect of low power on future research in the field of music education is the method in which non significant results are analyzed. If a study fails to uncover a

significant difference or a significant correlation, is the study necessarily a failure? In one instance, Keppel advised that a third possible decision in significance testing may be more practical; "suspending judgment," (Keppel, 1991, p. 181). This third possibility provides an entirely new light with which to interpret ambiguous results while simultaneously avoiding both Type I and Type II error, and the power analysis of the study would then provide the electricity in which to guide this new light. For example, a non significant study involving a high alpha and low power would be interpreted differently from another non significant study with a low alpha and high power. All non significant results are not created equal, and a power analysis can easily point out the difference.

A final consequence of low power on music education research is the danger inherent in Type I error. Realistically, any researcher designing a study without adequate attention to effect size and statistical power might be guilty of formulating a poorly conceived and ultimately invalid null hypothesis. Consequently, if the null hypothesis is invalid, then the door is wide open for Type I error. According to Cohen (1965), Type I error is four times more critical than Type II error, and the dangerous effects of spuriously positive results can never be ignored. In many cases, curriculum policy and funding decisions for public school music programs are made as a direct result of someone's quantitative study. To place the future of a child's musical development in the hands of a poorly and inadequately designed quantitative study is unthinkable.

If overall statistical power in music education is low, then what can be done to correct the problem? First, a higher alpha level could be used. Raising alpha would lower beta, which means a higher statistical power would result. However, a greater chance for Type I error would also result from a higher alpha. This solution is certainly as undesirable as the problem.

Second, music education research could attempt to uncover larger effects. Since the estimation of effect size is based primarily on an *a priori* review of the literature, then effect size seems tied to past research results. In other words, how can a researcher justify using a large effect size when similar past research seems to center around uncovering small effects? Besides being unethical, simply "to raise statistical power" is not enough of a justification to contradict past research results. Although the researcher is responsible for determining the effect size for his study, many are frightened away by an unfortunate ignorance of the topic. For this reason, the aid of having a small, medium, and large effect size value already computed can eliminate some of the confusion. These three pre-determined values of effect size should have the result of making power analysis more accessible to the average researcher.

Third, a researcher might be tempted to solve the problem of low power by simply saying, "Raise all sample sizes." Although this is certainly one very viable option, it is not a panacea for all that ails music education research. Realistically, part of the fault lies with the very nature of significance testing. If research that failed to uncover significance was omitted from publication in both the *JRME* and this study, then the statistical power of the overall work undertaken in the field of music education should be substantially lower than the median power reported for the *JRME*; the *JRME* only represents the successful research and not all of the research. Published research must certainly be more powerful than unpublished research. Since we cannot conclude that the unpublished work in the field had higher statistical power then we must conclude that the overall field of music education contains lower power than .64 for medium effects as uncovered in this study. Perhaps future studies could examine other research publications in the field of music, such as the *Southeastern Journal of Music Education*, the *Journal of Band Research*, the *Journal of Music Therapy*, the *Journal of Research in Singing*, and the *Psychology of Music* for comparison. Also, unpublished sources could be examined for power analysis, such as doctoral dissertations, conference papers, ERIC documents, etc.

The most practical solution to the problem of low statistical power seems to be to simultaneously raise the sample size and lower the number of independent variables. Power can be raised by simply raising the sample size. Yet, in articles using chi square, analysis of variance, multiple regression and correlation, and multivariate techniques, the statistical power could have been raised by trimming the extraneous variables. Since these statistical procedures depend on cell quantity and/or numerator and denominator degrees of freedom for their power calculations, then their manipulation would directly influence power. In every one of the above procedures, lowering the number of cells or the degrees of freedom has the direct result of slightly raising statistical power. This problem of controlling the influence of too many extraneous variables could simultaneously increase both statistical power and overall research design quality.

Based on these findings, many suggestions can be offered to improve the overall quality of research designs in music education. First, more attention should be placed on *a priori* power analyses of research designs. If power analyses are conducted *a priori*, they can be of the most benefit to music education research by uncovering flaws in the research design before the study is ever launched and time and money is ever spent. This would place more emphasis on the initial planning stages of the research project (where it should be) and reduce the number of studies begun both prematurely and with inadequate designs.

Second, adequate sample sizes should be used in order to achieve adequate power. This recommendation does not imply that all sample sizes should be increased dramatically,

because a study with too many subjects can lead to power that is too high and to mountains of trivial significance. According to the findings of this study, the sample sizes used in music education research are not inadequate at all. If proper attention is paid to *a priori* statistical power, then the sample sizes will be the exact number needed in order to maintain a power of .80 for the desired effect size.

Third, a greater understanding and application of the concept of effect size is needed in music education research. Any consumer of research (music education or otherwise) with an understanding of effect size can more accurately apply research results to personal and professional situations if that research includes estimates of effect size. Most researchers are seeking some kind of effect size in their studies whether they know it or not, and if some studies in a particular area of music education are seeking to detect small effects while others are seeking large effects, then the consumers deserve to know which effect size is targeted in the study in order to be able to more accurately comprehend those results. According to Cohen (1990), the proper way to interpret research results is through an examination of effect size and not a simple glance at the p value.

The only way that a music education researcher can be completely comfortable and confident with the results of a study is when the alpha level, effect size, sample size, and research design combine to yield a high statistical power. Only then will he or she know that the chances for Type I and Type II error have been effectively eliminated and the research is in an excellent position for a valid rejection of the null hypothesis. Therefore, this study should serve notice to music education investigators of both significant and non-significant results to take heart; perhaps all that was needed for a successful publication was an effective *a priori* power analysis. A research design created with statistical power in mind could lead directly to a successful conclusion. Although every study cannot possibly be published, larger samples and higher power can help lead to a better and ultimately successful manuscript.

Although this study tends to chastise, music educators should proudly note the final results of the power analysis. An overall median power of .64 for medium effects is quite a bit higher than the results found in many of the other similar studies of the other branches of the behavioral sciences. This study should serve as an inspiring testimony to music educators as to how far music education research has actually come since the founding of the *JRME* in 1953. Every performing musician understands that a near flawless recital will sometimes be remembered for that one sour note. Perhaps, in a reverse analogy, this study could be remembered for that one good note; the fact that music education researchers currently publishing in the *JRME* have much room for improvement and much more for which to be proud.

REFERENCES

Borenstein, M. & Cohen, J. (1988). *Statistical Power Analysis: A Computer Program.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Daniel, T. D. (1992). Observations of the statistical procedures used in the *Journal of Research in Music Education.* Paper presented at the meeting of the Mid-South Educational Research Association, Knoxville, TN.

Kendall, M. J. (1988). Two instructional approaches to the development of aural and instrumental performance skills. *Journal of Research in Music Education, 36*, 205-219.

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

Kratus, J. (1989). A time analysis of the compositional processes used by children ages 7 to 11. *Journal of Research in Music Education, 37*, 5-20.

Mark, M. L. (1986). *Contemporary Music Education* (2nd ed.). New York: Schirmer.

Neyman, J., & Pearson, E. S. (1928a). On the use of interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A*, 175-240.

Neyman, J., & Pearson, E. S. (1928b). On the use of interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika, 20A*, 263-294.

Radocy, R. E. (1988). Forum. *Journal of Research in Music Education, 36*, 204.

Rainbow, E. L., & Froehlich, H. C. (1987). *Research in Music Education: An Introduction to Systematic Inquiry*. New York: Schirmer.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years?" *Journal of Consulting and Clinical Psychology, 58*, 646-656.

Schmidt, C. P. (1989). Applied music teaching behavior as a function of selected personality variables. *Journal of Research in Music Education, 37*, 258-271.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Shavelson, R. J. (1988). *Statistical Reasoning for the Behavioral Sciences*. Needham Heights, Massachusetts: Allyn and Bacon.

Table 1

Median Sample Sizes of JRME Articles, 1987-1991

	True sample size	Optimum sample size *	Total number of statistical tests
1987	122	164	27
1988	116	157	51
1989	75	161	52
1990	97	141	60
1991	114	130	65
Overall	99	145	255

* Optimum sample size refers to alpha=.05, medium effect size, and power=.80.

Table 2
Estimated Effect Sizes by Procedure (Cohen, 1988)

Statistical procedure	Small effect size	Medium effect size	Large effect size
t Test for means	d=.2	d=.5	d=.8
Significance of a product moment	r=.10	r=.30	r=.50
Differences between correlation coefficients	q=.10	q=.30	q=.50
Test that a proportion is .50 and the sign test	g=.05	g=.15	g=.25
Differences between proportions	h=.20	h=.50	h=.80
Chi square	w=.10	w=.30	w=.50
ANOVA/ANCOVA	f=.10	f=.25	f=.40
Multiple regression and correlation	f^2=.02	f^2=.15	f^2=.35
Set correlation and multivariate methods	f^2=.02	f^2=.15	f^2=.35

Effect Size Index Legend (Cohen, 1988)

d=Effect size index for *t*-tests of means expressed in standard units.
r=Effect size index expressed as a population correlation coefficient.
q=Effect size index expressed as a Fisher **z** transformation of **r**.
g=Effect size index expressed as the distance in units of proportion from .50.
h=Effect size index expressed as a nonlinear arcsine transformation of population proportions.
w=Effect size index expressed as the discrepancy between paired proportions over cells.
f=Effect size index expressed as the standard deviation of the standardized means.
f^2=Effect size index expressed as the squared standard deviation of the standardized means.

Table 3A
Overall Median Statistical Power of the JRME, 1987-1991

Power	Small effects		Medium effects		Large effects	
	Frequency	Cumulative percentage	Frequency	Cumulative percentage	Frequency	Cumulative percentage
.99 +	1	100	20	100	112	100
.95-.98	0	99.6	14	92.2	25	56.1
.90-.94	0	99.6	25	86.7	35	46.3
.80-.89	0	99.6	33	76.9	38	32.5
.70-.79	0	99.6	18	63.9	17	17.6
.60-.69	1	99.6	26	56.9	11	11.0
.50-.59	2	99.2	30	46.7	11	6.7
.40-.49	1	98.4	40	34.9	4	2.4
.30-.39	15	98.0	24	19.2	2	0.8
.20-.29	26	92.2	17	9.8	0	0
.10-.19	145	82.0	8	3.1	0	0
.03-.09	64	25.1	0	0	0	0
Total *N*	255	-	255	-	255	-

	Small effects	Medium effects	Large effects
Mean	0.15	0.64	0.90
Median	0.13	0.64	0.97
Mode	0.11	0.47	1.00
Std dev	0.10	0.25	0.14

Table 3B
Median Statistical Power by Volume

	Small effects	Medium effects	Large effects	Total *N*
1987	0.18	0.86	1.00	27
1988	0.12	0.66	0.99	51
1989	0.11	0.50	0.905	52
1990	0.125	0.65	0.98	60
1991	0.13	0.66	0.97	65

Table 4A
Statistical Power of Analysis of Variance

Power	Small effects		Medium effects		Large effects	
	Frequency	Cumulative percentage	Frequency	Cumulative percentage	Frequency	Cumulative percentage
99 +	1	100	5	100	36	100
.95-.98	0	99.2	2	96.1	13	71.7
.90-.94	0	99.2	11	94.5	26	61.4
.80-.89	0	99.2	14	85.8	28	40.9
.70-.79	0	99.2	5	74.8	12	18.9
.60-.69	1	99.2	12	70.9	5	9.4
.50-.59	1	98.4	23	61.4	7	5.5
.40-.49	0	97.6	31	43.3	0	0
.30-.39	9	97.6	14	18.9	0	0
.20-.29	7	90.6	10	7.9	0	0
.10-.19	75	85.0	0	0	0	0
.06-.09	33	26.0	0	0	0	0
Total *N*	127	-	127	-	127	-

	Small effects	Medium effects	Large effects
Mean	0.15	0.584	0.885
Median	0.11	0.51	0.91
Mode	0.11	0.47	1.00
Std dev	0.114	0.222	0.121

Table 4B
Median Statistical Power of ANOVA Procedures

	Small effects	Medium effects	Large effects	Total *N*
One-way	.125	.60	.96	28
Factorial	.11	.50	.905	90
ANCOVA	.11	.56	.96	9

Table 5A
Statistical Power of Multivariate Procedures

Power	Small effects		Medium effects		Large effects	
	Frequency	Cumulative percentage	Frequency	Cumulative percentage	Frequency	Cumulative percentage
.99 +	0	100	3	100	14	100
.95-.98	0	100	4	83.3	0	22.2
.90-.94	0	100	0	61.1	0	22.2
.80-.89	0	100	4	61.1	1	22.2
.70-.79	0	100	3	38.9	1	16.7
.60-.69	0	100	0	22.2	0	11.1
.50-.59	0	100	0	22.2	1	11.1
.40-.49	0	100	0	22.2	0	5.6
.30-.39	2	100	2	22.2	1	5.6
.20-.29	4	88.9	1	11.1	0	0
.10-.19	9	66.7	1	5.6	0	0
.03-.09	3	16.7	0	0	0	0
Total *N*	18	-	18	-	18	-

	Small effects	Medium effects	Large effects
Mean	0.162	0.762	0.915
Median	0.15	0.87	1.00
Mode	0.13	0.87	1.00
Std dev	0.084	0.277	0.184

Table 5B
Median Statistical Power of Multivariate Procedures

	Small effects	Medium effects	Large effects	Total *N*
MANOVA	.11	.72	1.00	7
Disc. analysis	.17	.92	1.00	6
MANCOVA	.185	.815	1.00	2
Canonical corr.	.18	.95	1.00	3

Table 6A
Median Statistical Power by Degree of Difficulty

	Small effects	Medium effects	Large effects	Total *N*
Basic	.13	.68	.99	107
Intermediate	.11	.52	.91	109
Advanced	.15	.87	1.00	18
Other	.15	.67	.98	21

Table 6B
Median Statistical Power Comparison of All Procedures

	Small effects	Medium effects	Large effects
One-way ANOVA	.125	.60	.96
Factorial ANOVA	.11	.50	.905
ANCOVA	.11	.56	.96
Chi-square	.12	.73	.99
Spearman rank	.15	.67	.97
Kruskal-Wallis	.14	.67	.97
Wilcoxon	.13	.53	.90
Mann-Whitney	.12	.49	.87
Friedman	.155	.89	1.00
Kendall	.07	.32	.73
MANOVA	.11	.72	1.00
Disc. analysis	.17	.92	1.00
MANCOVA	.185	.815	1.00
Canonical corr.	.18	.95	1.00
t-test	.145	.625	.955
Multiple regression	.185	.92	1.00
Correlation	.17	.87	1.00
z-test	.09	.385	.68