EMBRY-RIDDLE
Aeronautical University
SCHOLARLY COMMONS

Department of Graduate Studies - Worldwide | College of Aeronautics

Summer 2011

An Analysis of Statistical Power in Aviation Research

David Carl Ison
Embry-Riddle Aeronautical University, isond46@erau.edu

Follow this and additional works at: http://commons.erau.edu/ww-graduate-studies

 Part of the Aviation Commons

Scholarly Commons Citation
Ison, D. C. (2011). An Analysis of Statistical Power in Aviation Research. *International Journal of Applied Aviation Studies, 11*(1). Retrieved from http://commons.erau.edu/ww-graduate-studies/11

This Article is brought to you for free and open access by the College of Aeronautics at Scholarly Commons. It has been accepted for inclusion in Department of Graduate Studies - Worldwide by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

An Analysis of Statistical Power in Aviation Research

David Carl Ison

Rocky Mountain College
1511 Poly Drive
Billings, MT 59102
406-657-1061
isond@rocky.edu

Abstract

This study sought to evaluate the statistical power of aviation research published in four prominent peer-reviewed journals (*Collegiate Aviation Review, Journal of Air Transportation Worldwide, Journal of Aviation/Aerospace Education and Research, and International Journal of Applied Aviation Studies*). Further, this study investigated whether power was mentioned or calculated as well as if articles included details on effect size(s). The study yielded 128 articles that included statistical testing and provided enough information to calculate power. From these articles a total of 1,692 statistical tests were analyzed. The average power of these tests was .277 considering a small effect size, .685 when considering a medium effect size, and .874 when assuming a large effect size. Considering that a medium effect size is generally utilized when there is no research-based reason to use an alternative level and that the accepted minimum power value is .80, aviation research appears to be underpowered. Also, only 5.6% of articles conducted an *a priori* power analysis whilst 11.9% mentioned power. Among studies that included statistical testing, only 4.2 % calculated effect size. Thus aviation research commonly fails to provide critical research data. Guidance on ways researchers can improve power and/or reduce sample size requirements are provided. Suggestions for future research and policies are also provided.

An Analysis of Statistical Power in Aviation Research

Null hypothesis significance testing (NHST) dominates the focus of research studies in a variety of fields with aviation being no exception (Borkowski, Welsh, & Zhang, 2001; Ferrin et al., 2007; Jones & Sommerlund, 2007). NHST investigates research problems by determining which of two alternatives – the first that there *is* a difference between groups (termed the alternative hypothesis or H_1), or the second, that there is *no* difference between groups (termed the null hypothesis or H_0) – is apparently true (Jones & Sommerlund, 2007; Stevens, 2007). Ferrin et al. (2007) described this model as one in which a researcher "calculates the test statistic, and if it is sufficiently large and the *p*-value is sufficiently small, the null hypothesis is rejected and the corresponding alternative hypothesis is accepted" (p. 87). This method of inquiry arose from the efforts of Neyman and Pearson in the early 20th century and has been widely adopted since (Cohen, 1992; Sedlmeier & Gigerenzer, 1989; Spanos, 1999). Not surprisingly, researchers put forth a tremendous amount of effort to seek statistical significance of a certain level in order to claim a difference, or lack thereof, between or among groups. The generally accepted norm for statistical significance is $\alpha = 0.05$ (Coladarci, Cobb, Minium, & Clarke, 2007; Stevens, 2007).

Even in light of its prevalence in the research literature, there are noteworthy concerns about

the appropriateness and utility of NHST. Fagley (1985) noted that if researchers were to ardently adhere to a veritable definition of the null hypothesis, it would always be determined to be false. Kline (2004) also noted that there are many fallacies within the literature about *p* values being equated to effect sizes and the false assumptions that if the null hypothesis is not rejected then it has to be true. Also, Kline (2004) displayed concern that only when the null-hypothesis is rejected are the findings considered of value to the research community.

Fisher (1966) disagreed with an *a priori* determination of a significance level (α), instead advocating the use of a sliding scale of significance proportionate to the *p*-value resultant from the conducted research. Cohen (1992) found that in most studies involving statistical tests, "the chance of obtaining a significant result was about that of tossing a head with a fair coin" (p. 155). Along the same lines, Ferrin et al. (2007) remarked that "unfortunately, knowing the *p*-value reveals nothing about either the magnitude of the effect or about the width of the interval on the distribution line (confidence interval), or about power; nor does it provide information about the practical or clinical significance of the finding" (pp. 87-88). It is not uncommon that details such as effect sizes, which are arguably just as important as *p*-values, if not more so, are regularly missing from research findings (Osborne, 2008).

Another problem that has been noted concerning archetypical significance testing is its focus on avoiding a Type I error, i.e. the rejection of a null hypothesis when in fact it is true (Cohen, 1962; Stevens, 2007). This concentration on the probability of performing a Type I error (α) often leads to the neglect of Type II (β) error avoidance. This oversight may lead to researchers having an undesirable chance of accepting a null hypothesis that is instead actually false. Simply, the probability of a study successfully detecting a difference among groups in order to reject a null hypothesis, known as power, is often very low. What is especially problematic about the prevalence of studies with low power is that these blunders can be easily avoided by conducting a power analysis during the research design process. Further, the findings of research can be scrutinized in terms of the actual power, i.e. studies that report "insignificant" findings but are determined to have low power should be viewed with skepticism (Ferrin et al., 2007).

Cohen (1962) first reported his concerns that "the problem of power is occasionally approached indirectly" and studies overwhelmingly pay "careful attention to issues of significance, and typically no attention to power" (p. 145). Kosciulek and Szymanski (1993) recognized similar deficiencies in research noting that "statistical power analysis is a desirable and necessary ingredient in planning and conducting effective research. Unfortunately, however, it is an underused tool in […] research" (p. 212). Over the last 50 years, there has been little improvement in the inclusion of power analysis in research. Investigations into studies conducted in areas such as psychology, medicine, behavioral accounting, business, and education found a large percentage had low power values or neglected power entirely (Aguinis, Beaty, Boik, & Pierce, 2005; Borkowski, Welsh, & Zhang, 2001; Ferrin et al., 2007; Jones & Sommerlund, 2007; Osborne, 2008). The absence of power testing raised concerns at the American Psychological Association (APA) which convened a Task Force on Statistical Inference which defined "guidelines indicative of good research" which included "the reporting of effect size estimates and confidence intervals for any effect size involving principal outcomes as well as consideration of statistical power and sample size in the design of studies" (Ferrin et al., 2007, p 88). The *Publication Manual of the American Psychological Association* (6th ed.) clearly notes that researchers should "provide evidence the study has sufficient power to detect effects of substantial interest" (APA, 2010, p. 30).

In light of the practical and statistical importance of power analysis, it is critical that research inquiries include such data. Several studies have been conducted in a variety of subject areas in efforts to determine the level of inclusion of power analysis to help shed light on the general quality of research and statistical analysis that exists in

a body of research. As aviation research has continued to expand and become more mainstream, it becomes ever more critical that it comply with general research standards, but what is even more essential is that the research being published provides meaningful and well-founded findings determined by competent research and analysis methods. Therefore this study analyzed the statistical power of quantitative aviation research studies found within four prominent aviation-related peer-reviewed academic journals – the *Collegiate Aviation Review*, the *Journal of Air Transportation World Wide*, the *Journal of Aviation/Aerospace Education and Research*, and the *International Journal of Applied Aviation Studies*. Two related publications, the *Journal of Aviation Management and Education* and the *International Journal of Professional Aviation Training Testing Researc*h, were omitted as these journals had a very limited quantity of articles to analyze.

Statistical Power

The power of a statistical test is defined as "the probability, given that H_0 is false, of obtaining sample results that will lead to the rejection of H_0" (Coladarci, Cobb, Minium, & Clarke, 2008, p. 403). More simply, power refers to the chance of a statistical test to detect a difference between or among groups being analyzed. Discussions about power normally mention the Type II error (β), which is the "probability of retaining the null hypothesis when it is false" (Coladarci, Cobb, Minium, & Clarke, 2008, p. 404) therefore power can be determined by the formula 1 – β. The resultant number can be viewed as the percent chance that the statistical test will be able to rightfully reject a false null hypothesis, e.g. a power of 0.33 means that the test has a 33% chance of succeeding to reject a false null hypothesis. Obviously, a study that only has a 33% chance at success is not very viable nor would one want to take findings of a study with such a level of power too seriously (Cohen, 1992; Ferrin et al., 2007).

Determinants of Statistical Power

Statistical power is most easily defined by the formula 1 – β, however, there are several additional factors that are involved in the calculation of power. There are five determinants of power: significance level, homogeneity of samples, sample size, effect size, and directionality. The significance level, or alpha (α), is the probability of rejecting a true null hypothesis (Type I error). This is commonly set at 0.05 meaning there is a 5% chance of committing a Type I error. Some studies go as far as using a higher α standard such as 0.01. Yet it is important to recognize the relationship between α and β. When a researcher demands a more stringent α, they simultaneously allow for a larger chance of committing a Type II error (β) (Stevens, 2007). Therefore Cohen (1988) suggested weighing the importance of α versus β during the research design process vis-à-vis arbitrarily setting $\alpha = 0.05$. The recommended procedure is to divide β by α to determine a ratio that ideally does not exceed 4 : 1. For example, if $\alpha = 0.05$ and $\beta = 0.20$, the resultant ratio would be 4 : 1. The power in this case would of course be 0.80 (1 – 0.20), i.e. there would be an 80% chance that the study would be able to correctly identify a difference among investigated groups. In sum, as α is strengthened, power is reduced, therefore it is no surprise that Stevens (2007) stated that "it is not always wise to set α as low as 0.05 or 0.01." (p. 105). Refer to Figures 1 and 2 (page 70) for a comparison of power when $\alpha = 0.05$ versus $\alpha = 0.01$.

Another factor in determining the power of a statistical test is the reliability or homogeneity of samples which can be observed through the standard error of a statistic ($SE_{\bar{x}}$) which is defined by a relationship between the population variance estimate (s^2) and the sample size (n) (Cohen, 1988):

$$SE_{\bar{x}} = \sqrt{s^2/n}$$

As is obvious with a constant sample size, a reduction in variance nets a lower standard error. The standard error of tests utilizing dependent samples is lower than if independent samples are utilized. This is due to the fact that "the standard error of the difference between means is modified to take

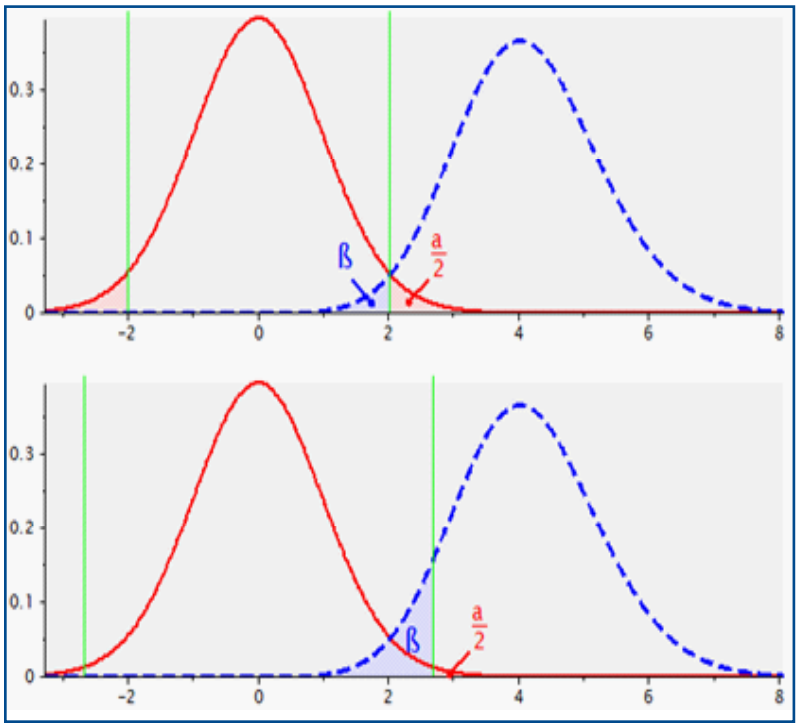

Figure 1. Comparison of Power (1 – β) between α = 0.05 (top) and α = 0.01 (bottom). Created in *G*Power.* *Note:* Power is indicated by the un-shaded region underneath the dashed curve.

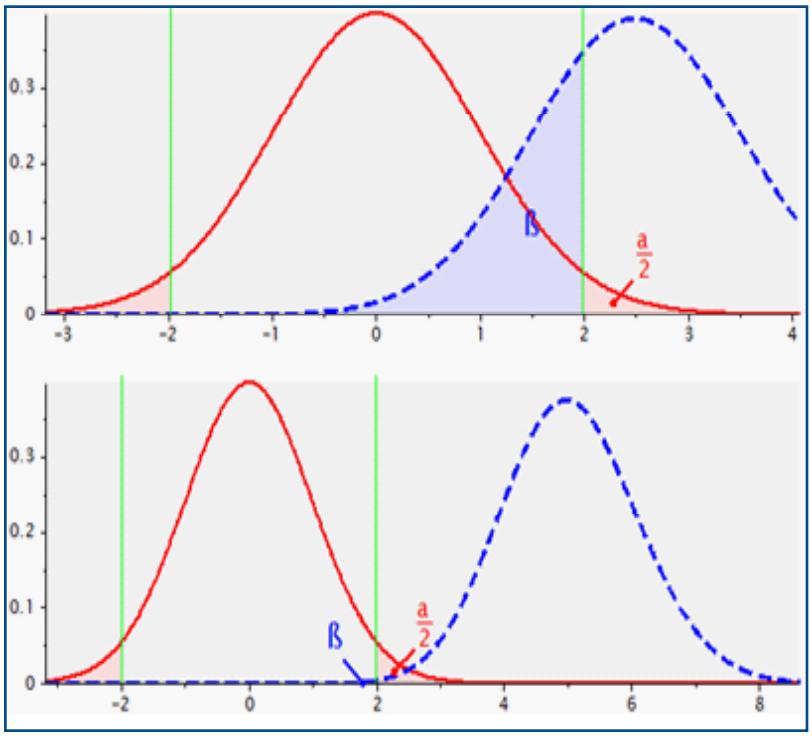

Figure 2. Comparison of Power (1 – β) between a *t*-test utilizing independent samples (top) and dependent samples (bottom). Created in *G*Power.* *Note:* Power is indicated by the un-shaded region underneath the dashed curve.

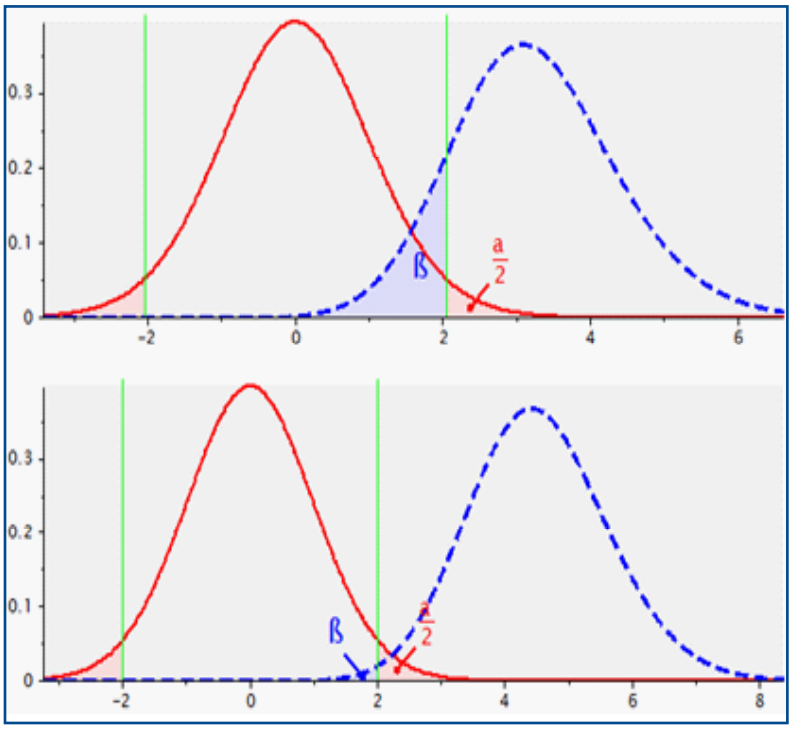

Figure 3. Comparison of Power (1 – β) between n = 30 (top) and n = 60 (bottom). Created in *G*Power*. *Note:* Power is indicated by the un-shaded region underneath the dashed curve.

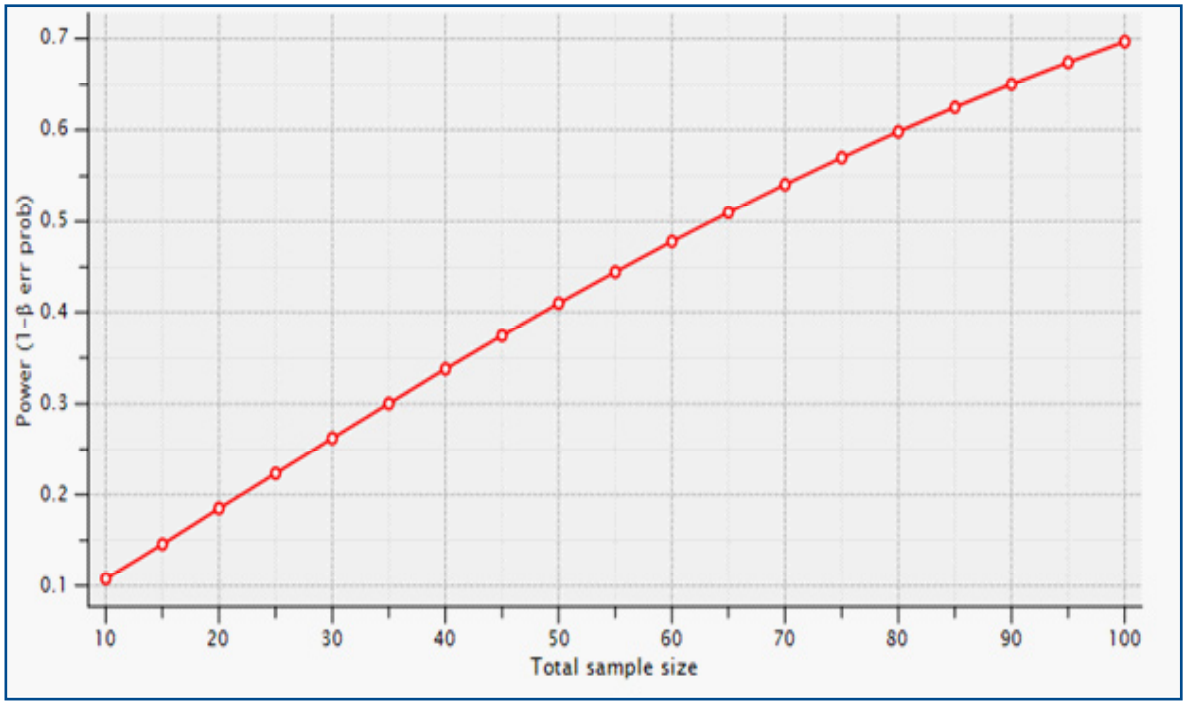

Figure 4. Plot of Power vs. Sample Size for an Independent Means t-test. Created in *G*Power*.

into account the degree of correlation between the paired scores" (Coladarci, Cobb, Minium, & Clarke, 2007, pp. 310-311). If standard error is reduced, the result is an increase in power. The use of dependent or homogeneous samples results in larger power value (Coladarci, Cobb, Minium, & Clarke, 2007).

Also, as the aforementioned formula indicates, as sample size increases the standard error would also be reduced. Therefore considering a constant variance, power increases with an increase in sample size (see Figures 3 and 4). An example of the influence of sample size on power can be seen if one utilizes a *t*-test for independent means. Keeping all other factors constant (two tailed, $\alpha = 0.05$, effect size of 0.50), the power of a study with n = 30 in each group would be 0.47 whilst if n were increased to 100 in each group power would grow to 0.94 (Cohen, 1988).

Another manipulator of power is effect size which Stevens (2007) defines as "how much of a difference the treatments make, or the extent to which the groups differ in the population on the dependent variable" (p. 106). Alternatively, Cohen (1988) defines effect size "as an index of degree of departure from the null hypothesis" (p. 10). Mathematically, effect size (δ) is calculated by dividing the difference between the means of investigated populations divided by the population standard deviation and is represented by the formula (Coladarci, Cobb, Minium, & Clarke, 2007):

$$\delta = \frac{(\mu 1 - \mu 2)}{\sigma}$$

When all other factors remain constant, as effect size increases power also increases (see Figure 5).

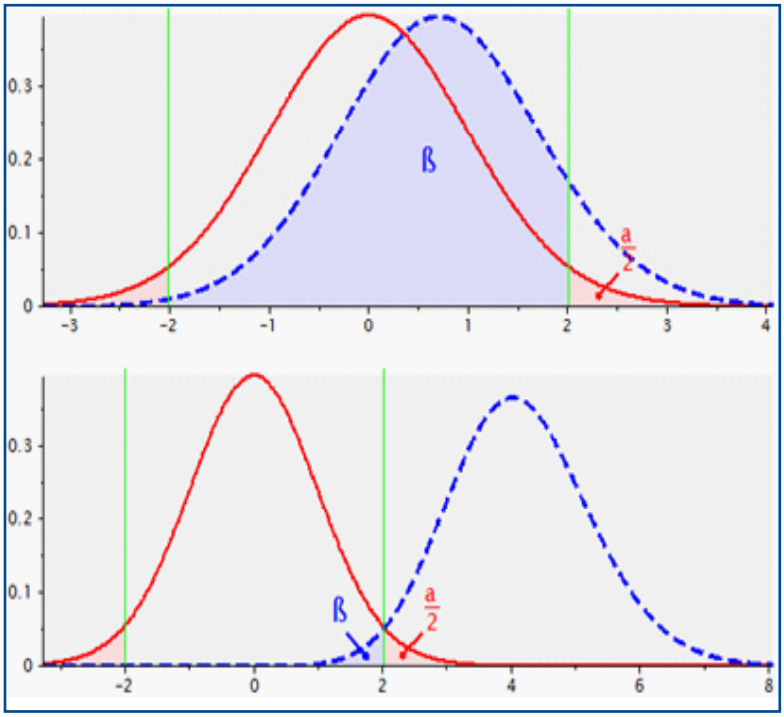

Figure 5. Comparison of Power (1 – β) between small d (top) and large d (bottom). Created in *G*Power. Note:* Power is indicated by the un-shaded region underneath the dashed curve.

This is due to the fact that the presence of a larger difference among groups would, in theory, be easier to detect (Cohen, 1988). The problem resides in the fact that "effect size is rarely known in advance" (Borkowski, Welsh, and Zhang, 2001). To assist in the selection of an effect size to use in power analysis, three general categories have been adopted: small, medium, and large. Cohen (1988) stated that:

> 'small' effect sizes must not be so small that seeking them amidst the inevitable operation of measurement and experimental bias and lack of fidelity be a bootless task [… and] large effects must not be defined so large that their quest by statistical methods is wholly a labor of supererogation (p. 13).

In most cases, it is logical to select "medium" effect so as to avoid one extreme or another. As Cohen (1988) described, medium effects would be perceptible to the naked eye. But because certain statistical test yield different levels of accuracy, individual tests have different δ values equating to designations of small, medium, and large. Effect sizes for common statistical tests are given in Table 1.

An example of the influence of effect size is if a *t*-test is performed with independent means, α = 0.05 and n = 100 in each group (note that effect size in *t*-tests is referred to as "d") (Cohen, 1992). If the researcher used a small d (0.20), the resultant power is 0.29. In contrast, if the recommended medium d (0.50) were used, the resultant power would be 0.94. When performing statistical analysis, researchers can select a one or two-tailed measure. If the researcher proposes a one-tailed measure and correctly identifies the directionality of the hypothesis, the critical area will be larger thus there is a higher likelihood that the null hypothesis will be rejected. As such, when all other factors remain constant, a one-tailed test will have a greater power than a two-tailed version (see Figure 6, page 74) (Coladarci, Cobb, Minimum, & Clarke, 2007). This advantage only exists, however, if the researcher surmises the correct direction (Cohen, 1988). The difference in power between a one-tailed and a two-tailed *t*-test of independent means (α = 0.05, n = 50 in each group, and d = 0.50) is 0.79 and 0.69 respectively.

Table 1.

Type of Statistical Test and Associated Acceptable Effect Sizes.

	Effect Sizes		
Test Type	Small	Medium	Large
1. *t*-test (independent means)	0.20	0.50	0.80
2. *t*-test (product-moment correlation)	0.10	0.30	0.50
3. Difference between two *r* values	0.10	0.30	0.50
4. Test vs. population proportion (P) = 0.50	0.05	0.15	0.25
5. Chi square – goodness of fit	0.10	0.30	0.50
6. One way ANOVA	0.10	0.25	0.40
7. Multiple correlation	0.02	0.15	0.35

Note: Adopted from Cohen (1992).

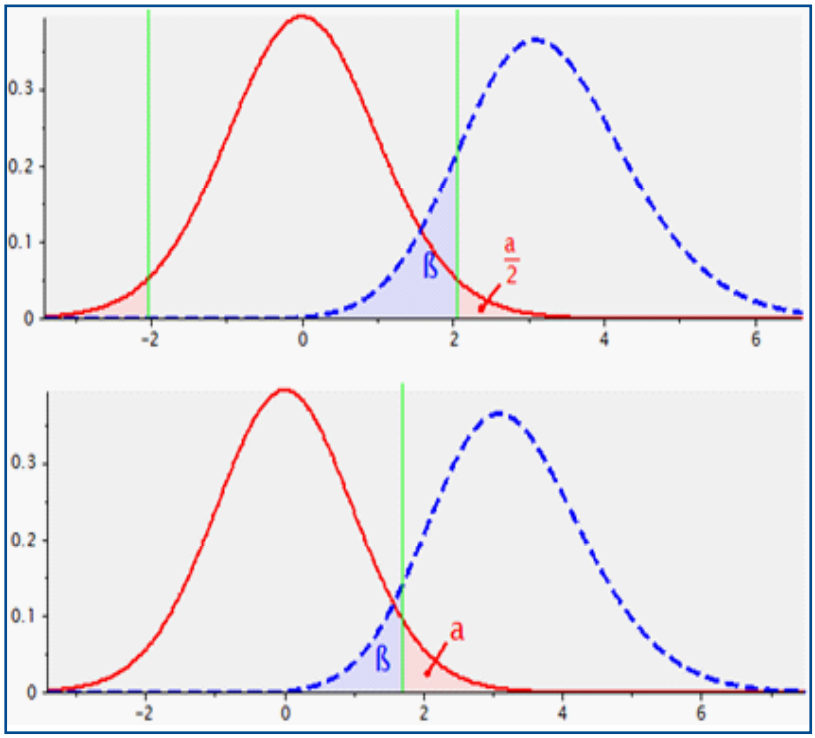

Figure 6. Comparison of Power (1 – β) between a two-tailed test (top) and a one-tailed test (bottom). Created in *G*Power*. *Note:* Power is indicated by the un-shaded region underneath the dashed curve.

Uses of Power: Incorporating Power into Research Design and Evaluation

There are two primary instances when statistical power analysis can be used in research – *a priori* and *a posteriori*. Ideally, researchers conduct a power analysis before partaking in their study so as to insure a reasonable chance of correctly rejecting a null hypothesis (Osborne, 2008). Cohen (1992) stated that a power of 0.80 or greater is acceptable. It is logical to perform this important step in research design because if a researcher determines that the power of the proposed study falls below 0.80, an amendment is in order to correct the deficiency. A common *a priori* use of power is the determination of sample size. Clearly researchers should determine the minimum number of participants in a particular study in order to have adequate power. At the same time, it may be advantageous to determine that fewer individuals are necessary to sufficiently undertake a study with a minimum power of 0.80 potentially saving the researcher time, money, and effort (Osborne, 2008).

Kosciulek and Szymanski (1993) outlined a pre-test power analysis plan that should be utilized by researchers during their methodology design process. The first step is to evaluate the literature to determine a reasonable effect size that can be expected when dealing with the subject at hand and the proposed experimental design. Next, the researcher should select an appropriate statistical test. With this information, the researcher can use power tables or statistical analysis software to determine the required sample size. The researcher can then estimate the power of the study. If the power is determined to be at or above 0.80, then the researcher can confidently move forward. If the power is below the desired level, the researcher can re-evaluate the sample size, alpha level, the proposed statistical test, or other aspects of the methodology for possible revision (Borkowski, Welsh, & Zhang, 2001; Kosciulek & Szymanski, 1993).

It is important to note that if a study uncovers statistically significant findings, either the study must have had sufficient power or a Type I error occurred. While this is true, it is important to consider that if the researcher in this case did not conduct an *a priori* power analysis, they were essentially blindly seeking results without any idea how likely they may be to find it, which is clearly an attribute of poorly designed research. Also, the consideration and inclusion of essential aspects related to power, such as effect size, are still critical to the presentation and analysis of findings (Cohen 1992; Kline, 2004; Osborne, 2008).

A posteriori approaches to power allow for more of an evaluation of the quality of research findings by peers. If a *post hoc* power analysis reveals low power in a study in which the null hypothesis was not rejected, "it is unclear whether a Type II error has occurred" (Osborne, 2008, p. 153). Equally, if a study that fails to reject the null hypothesis is revealed to have power of 0.80 or greater, readers can have a confidence that the study came to right conclusion (Osborne, 2008).

Previous Studies on Power Analysis in Research

Because of the crucial importance of adequate power among studies, there has been an assortment of research that has analyzed literature in an array of fields. The seminal study of power in research literature was conducted by Cohen (1962) in which 78 articles in Volume 61 of the *Journal of Abnormal and Social Psychology* were examined. Eight articles were found to be missing statistical testing and were omitted. Cohen (1962) then calculated power for each of the remaining articles. When considering small effect sizes, the mean calculated power among the studies was 0.18. "When one posits medium effects in the population (generally of the order of twice as large as small effects) the studies average[d] slightly less than a 50-50 chance of successfully rejecting their major null hypothesis" (Cohen, 1962, p. 150). When calculated assuming a large effect, the mean power rose to 0.83. Considering that "in the absence of any basis for specifying an alternative to the null hypothesis for purpose of power analysis, the criterion values for a medium effect are […] convention" (Cohen, 1962, p. 153) and the minimum power deemed acceptable by Cohen (1962; 1988; 1992) is 0.80, the reviewed research fell well short of the desirable power levels.

Sedlmeier and Gigerenzer (1989), using the work of Cohen (1962; 1988) as a model, investigated the power of a much broader range of journals in subject areas including psychology, education, communication, sociology, forensics, speech and hearing, communications, journalism, and marketing. When viewed with the assumption of a small effect size, only one journal had a mean power above 0.50. With a medium effect, two journals mean powers above 0.80 with more than half concentrated around the 0.50 mark. Even when considering large effects, five groups of journals did not meet the recommended power threshold of 0.80. Sedlmeier and Gigerenzer (1989) also analyzed 56 articles for their inclusion of power and discussions of why significance levels and sample sizes were selected. Only two mentioned power and in only four articles "alpha was mentioned, either by saying that it was set at a certain level (0.05) before the experiment or by referring to the danger of alpha inflation" (p. 311). No articles were found to include reasoning behind why a particular alpha levels or sample sizes were utilized.

Kosciulek and Szymanski (1993) examined 150 rehabilitation counseling studies containing 32 statistical tests. Within this literature, it was discovered that:

> 100% of the studies did not have a 50-50 chance of detecting small effect sizes. Furthermore, only 12 had a 1 in 2 change of finding significant results assuming medium effects. A comparatively small 9% of the studies showed less than a 50-50 chance of detecting large effects, and a miniscule 3% showed less than 3 in 10 chances (p. 212).

A study of accounting related literature was conducted by Borkowski, Welsh, and Zhang (2001) and included articles from three journals over a period between 1993 and 1997. In total, 258 articles with over 14,000 statistical tests within them

were analyzed. The average power among all journals over the five year period evaluated was 0.23 considering a small effect size, 0.71 when using a medium effect size, and 0.93 for large effect size.

Bezeau and Graves (2001) found slightly more encouraging results through a scrutiny of 66 clinical neuropsychology studies among three journals between 1998 and 1999. It was found that the mean power for studies assuming a 0.50 effect size to be 0.50, with those at the 0.80 effect size power was 0.768, and for those with an effect size of 1.35, the mean power was 0.957. Yet this study identified general deficiencies in statistical methods that were used noting that "few studies appear[ed] to conduct *a priori* power analyses; only 3% of the reviewed studies reported such an analysis [… and] only 9% of the reviewed […] studies *explicitly* reported the effect size of their results" (Bezeau & Graves, 2001, p. 403).

The plethora of research supporting the calculation of power prompted Osborne (2008) to attempt to identify if the inclusion of such statistical analysis has improved over time. The power values discovered by Cohen (1962) were compared to 96 educational psychology journal articles from 1998-1999. The findings indicated "significant but modest differences in observed power" (Osborne, 2008, p. 156) however a majority of articles still failed to surpass the desirable 0.80 power level. Among the more recent articles, the mean power presuming a small effect was 0.27, with a medium effect it was 0.71, and with a large effect it was 0.89. Only 2% of articles in the study discussed power and only 16.7% reported effect size.

Method

The journals included in this study were selected as they are representative of the research being conducted on subjects specific to aviation. This study includes the *Collegiate Aviation Review,* the *Journal of Air Transportation World Wide*, the *Journal of Aviation/Aerospace Education and Research*, and the *International Journal of Applied Aviation Studies*. Two related publications, the *Journal of Aviation Management and Education* and the *International Journal of Professional Aviation Training Testing Research*, were omitted as there were too few articles in each from which to make meaningful conclusions. The date ranges of the journal issues that were included in this study are listed in Table 2. These journals yielded 459 research articles. Each of these articles was carefully examined to determine whether or not they contained any type of statistical tests. All types of inferential statistics were included, e.g. parametric analyses such as tests of mean differences, correlation, regression, etc. Non-parametric analyses, e.g. chi square, Mann Whitney U, etc., were also included. Further, if the article came in Adobe PDF, Microsoft Word, or other searchable text document, the keyword "statistic" was used to serve as a confirmation that all statistical data were detected.

G*Power 3.1 and PASS 2008 software were used to conduct a *post hoc* power analysis for each test identified within the included articles. This calculation was based upon the statistical test used, sample size, and alpha level provided in the article. Power analysis was conducted at small, medium,

Table 2.

Issues/Date Ranges of Included Journals.

Journal Name	Date Range
Collegiate Aviation Review	1985 – Spring 2010
Journal of Air Transportation World Wide	1996 – 2004
Journal of Aviation/Aerospace Education and Research	1990 – 2003a
International Journal of Applied Aviation Studies	2003 – Summer 2010

and large effect sizes as outlined by Cohen (1988; 1992) with the value of effect size being tailored for each specific type of statistical test that was conducted. Unless an article specifically noted that a one tailed test was conducted, power analyses were calculated assuming a two tailed test.

An example of the calculation process follows. Assume a study utilized a two tailed *t*-test to analyze the difference between two independent means. Within this study, the researcher selected an alpha level of 0.05 and had two independent samples both of which included 30 individuals. Using the guidance of Cohen (1992), power for the effect sizes of small (0.20), medium (0.50), and large (0.80) can each be evaluated. For a small effect size, power would be 0.118 and for a medium effect size, power would be 0.477. As a medium effect size is generally considered a reasonable level, this study would have poor power. In fact, there is less than a 50% chance that the study will correctly identify a difference between means if it exists. Only a study assuming a large effect size would have adequate power, in this case it would be 0.861.

Articles were also analyzed to determine if the authors had conducted an *a priori* power analysis. Further, each article was evaluated to establish whether or not power was mentioned or considered. Lastly, articles were assessed for the presence of effect size calculations. These three details were uncovered through a thorough reading of the article. Further, if the article came in Adobe PDF, Microsoft Word, or other searchable text document, the keywords "power" and "effect size" were used to serve as a confirmation that the appropriate measures were detected.

Results

The *Collegiate Aviation Review (CAR)* included 155 articles with 41 containing statistical analysis. As the data were analyzed, it was discovered that there were several articles that failed to provide enough detail to conduct a power analysis. Among the *CAR* articles with statistical tests, 6 (14.6% of articles having statistical tests) omitted key details resulting in 35 articles that allowed for power analyses. A total of 580 statistical tests were conducted within these studies with an average of 16.5 tests per article. Within the issues of the *Journal of Air Transportation World Wide (JATW)*, there were 104 articles of which 29 included statistical tests. In the *JATW* there were 4 (13.7%) articles in which power analyses were not possible leaving at total of 25 articles that could be utilized. In these remaining articles there were 463 tests with an average of 18.5 tests per article. The *Journal of Aviation/Aerospace Education and Research (JAAER)* contained 40 articles of which 7 included statistical testing. However, 1 (14.2%) article lacked sufficient data to calculate power, thus 6 articles were able to be analyzed leaving 29 overall statistical tests resulting in an average of 4.8 tests per article. The *International Journal of Applied Aviation Studies (IJAAS)* included 160 studies with 65 containing statistical data. Three (4.6%) articles in the *IJAAS* had inadequate data to examine power leaving 62 articles to be studied. Within these articles, there were 620 tests conducted with an average of 10.0 tests per article. Across the 4 journals included in this study, the total number of articles that included the necessary information to conduct power analyses was 128. Within these articles there were 1,692 statistical tests conducted (see Table 3).

Each article identified to have statistical tests within it was examined so as to extract the necessary information to calculate power. Next, power analyses were conducted at the small, medium, and large effect sizes for each identified statistical test. In all but a few limited cases, G*Power 3.1 was sufficient to calculate power. In the instances that G*Power was lacking an applicable calculation, PASS 2008 was utilized. In the limited number of cases in which neither software package offered a solution (e.g. for MANCOVA), per the recommendations of Cohen (1962) and Dattalo (2008), substitutions were made for tests that were calculable by available software. Such substitutions have the tendency to slightly overrate the power (Cohen, 1962). For each individual publication, all of the power analyses for each statistical test were averaged for the small, medium, and large

Table 3.

Summary of Articles and Statistical Tests Included in this Study

Journal Name	# Articles (%)	# Stat. Tests (%)
Collegiate Aviation Review	35 (27.3)	580 (34.3)
Journal of Air Transportation World Wide	25 (19.5)	463 (27.4)
Journal of Aviation/Aerospace Education and Research	6 (4.7)	29 (1.7)
International Journal of Applied Aviation Studies	62 (48.5)	620 (36.6)
Total Averages (All Journals)	128 (100)	1,692 (100)

Table 4.

Summary of Power Analyses per Each Level of Effect Size for Each Journal.

Journal Name	Small ES	Medium ES	Large ES
Collegiate Aviation Review	.156	.697	.915
Journal of Air Transportation World Wide	.428	.749	.906
Journal of Aviation/Aerospace Education and Research	.144	.410	.623
International Journal of Applied Aviation Studies	.274	.614	.796
Total Averages (All Journals)	.277	.685	.874

Table 5.

Percent of Articles Including A Priori Power Analysis, Mention of Power, and Mention of Effect Size.

Journal Name	*A Priori*	Power Mentioned	ES Mentioned
Collegiate Aviation Review	0.6%	1.3%	0.6%
Journal of Air Transportation World Wide	0.0%	0.0%	3.4%
Journal of Aviation/Aerospace Education and Research	2.5%	2.5%	2.5%
International Journal of Applied Aviation Studies	9.5%	22.2%	4.8%
Average % (All Journals)	5.6%	11.9%	4.2%

effect sizes. The results of these analyses are aggregated in Table 4.

Articles where then examined for the calculation of an *a priori* power analysis. Among the 41 articles in the *CAR* only 1 (0.6%) included such an analysis. Of the 29 *JATW* articles with statistical tests, none reported a power analysis. One (2.5%) of the 7 articles in the *JAAER* contained a power analysis while such was present in 6 (9.5%) out of 65 articles in the *IJAAS*. Upon assessing the articles for the inclusion of any type of discussion of statistical power it was found that 2 (1.3%) of *CAR* articles, zero of *JATW* articles, 1 (2.5%) of *JAAER* articles, and 14 (22.2%) of IJAAS articles mentioned power. Effect size was mentioned in 1 (0.6%) of *CAR* articles. Within the *JATW*, 1 (3.4%) article discussed effect size. The *JAAER* also had 1 (2.5%) article referencing effect size. Lastly, remarks about effect size were included in 3 (4.8%) *IJAAS* articles. A summary of these results is presented in Table 5.

Discussion

It is readily apparent that aviation research studies are often underpowered and neglect to provide critical components necessary to confirm the soundness of such studies. If one considers a small effect size, there was only a slightly better than a 1 in 4 chance of detecting a difference. Considering a medium effect size, the average power was .685 which is still short of the generally acceptable .80 value. Only if considering a large effect size, which it is important to note is "roughly twice as large as medium" (Cohen, 1962, p. 150), would researchers exceed the .80 threshold. What is more problematic is that so few studies actually considered power and among the studies that did mention power, the calculation thereof was rarely conducted. The neglect of effect size makes it more difficult for the research community to garner the true significance of a study by the lack of appropriately framing findings.

Some other related issues also arose during this research. Fourteen (9.8%) of the 142 articles that included statistical tests failed to provide enough information to conduct a *post hoc* power analysis. This was generally due to incomplete or missing sample data or omitted details concerning a statistical test (e.g. numbers of groups or degrees of freedom). Several articles did not cite the results of statistical tests in APA or any other recognizable format. Three articles stated that a particular statistical test was done and that the results were either significant or not, but no further details were provided such as the actual test statistic and associated elements. One article stated that statistical testing was done, but no specific test was mentioned. Further the article went on to state the findings were significant but yielded no additional information. Two studies claimed abnormally large effect sizes which naturally boosted the power of the study even in light of the use of small sample sizes. These studies cited that such effect sizes were chosen based on the findings of previous research. However, upon closer examination, the sample membership was dissimilar to the individuals studied in the cited research, therefore making the choice of effect size somewhat questionable.

These findings are problematic for several reasons. Much of the research examined in this study was underpowered when considering a medium effect size. This means that the studies had a less than acceptable likelihood of identifying a difference or effect if one was actually present. As aviation is such a safety sensitive industry, it is critical that related research be able to adequately identify what is sought and that key findings are not missed from poorly designed or conducted research. What is more troubling is that the studies in the examined journals are probably the highest powered studies conducted in these subject areas as Cohen (1962) noted "if anything, published studies are more powerful than those which do not reach publication, certainly not less powerful" (p. 152). Thus there is probably more research that is being conducted within the industry that has even lower power.

The infrequent inclusion of vital components such as power, fundamental to the establishment of an adequate sample size, effect size, and sound

statistical reporting is extremely disconcerting. As Spybrook (2008) stated:

> for reviewers to be able to confidently assess whether a study has adequate power, the parameters required to conduct a power analysis must be included […]. The failure to report these parameters causes two problems: (a) the reviewers cannot replicate the analysis and (b) the reviewers cannot judge the appropriateness of the parameters used in the analysis. (p. 230)

Again this disserves the aviation industry. While the lack of the mention of power analysis does not guarantee that it was not appropriately assessed, its omission leaves readers to wonder if the researcher did in fact consider it. The merit of research is directly related to the ability to reconstruct a particular study. Missing information calls the dependability of such research into question. Moreover, in order for the aviation industry to make improvements and gains in understanding, stakeholders need to be provided with sound, well-conceived research.

It is clear that aviation research is often underpowered and frequently underreports effect size and power however this should be kept in perspective. The performance of aviation research should be compared to other subject areas in recent research. The Borkowski, Welsh, and Zhang (2001) study of over 14,000 accounting articles yielded an average power of 0.71 with medium effect size. Osborne (2008) similarly found that educational psychology articles in 96 journals had the same average power, 0.71, at the medium effect size. Recall that the average power calculated in this study was .685 which is closely comparable. Bezeau and Graves (2001) found that 3% of neuropsychology articles that were examined mentioned power and 9% calculated effect size. Osborne (2008) discovered that 2% of educational psychology articles in the study mentioned power while 16.7% reported effect size. This study found that aviation research mentioned power in 11.9% of articles and effect size was calculated in 4.2% of cases. So in the case of power, aviation research is at least performing better in recognizing this important aspect of statistical analysis, yet aviation research is apparently lagging in the reporting of effect sizes.

The question that remains is what can be done to improve future aviation research? Considering that most studies tend to use $\alpha = 0.05$ and, assuming a medium effect size (as many studies do not have a known or defined effect size), the problem appears to lie with sample size. As Cohen (1962) noted "if we then accept the diagnosis of general weakness of the studies, what treatment can be prescribed? Formally, at least, the answer is simple: increase sample sizes" (p. 151). Of course, there will be times when sample sizes are limited due to a variety of constraints, for example fiscal or practical limits. It is not uncommon for aviation program populations to be so small that extracting ample numbers for samples, particularly if multiple groups are required, is not possible.

Considering that small sample sizes are common in aviation research, lamenting the need to increase sample size is not practical and provides no solutions to aviation researchers. Instead, researchers need a toolbox to access during their research design process in order to maximize power even if it does not reach the minimums advocated in the literature.

One method to increase power is to accept a larger alpha level. In studies that do not have immediate safety or large financial implications, a higher tolerance for Type I errors could be accepted. Thus diversions from what is generally considered "the norm" may be viable options in certain situations. Leahey (2005) rigorously argued that blindly selecting the .05 significance level is problematic and the individual research setting should be considered when selecting alpha levels. There are instances when it is certainly reasonable to use a "non-standard" alpha of .10. According to the University of New England (2000), there are even cases where an alpha of .20 may be reasonable. Regardless of the choice of alpha, "at a minimum, the reporting of β would [help to] complement and interpret the true value of a reported α in any given study" (Cohen, 1962, p. 82). As is true with any well conducted study, all decisions in research design such as determining sample size, α, and β

levels should be backed with ample and appropriate citation support.

Another potential way to manage power and sample size is to further investigate or reconsider the effect size that is expected. Whilst it is often not possible to know what the effect size is going to be, it is worth digging into existing literature to see if anything similar has been done in the area of interest. If a larger effect size can be used in the power calculation, a smaller sample size or lower power would be required.

Researchers can also consider the use of a one-tailed test in lieu of a two-tailed test. Again, this choice should be supported by evidence in the literature or if a critical component of the proposed inquiry. If a researcher can justify that there is an inclination for a directional hypothesis, e.g. looking for an increase rather than simply a difference between groups, then they can gain power or take the advantage of lowering the required sample size.

Another way that researchers can reduce their sample size burden or boost power is to design the study using dependent samples. Because of the lower variance between these groups, researchers gain the aforementioned benefits. Clearly, not all studies lend themselves to be changed to this design, but it is worthy of consideration when pressed for power or sample size.

Researchers should be aware that different formulas are used in the calculation of power for each type of statistical test, therefore there is some variance in the power demands among individual tests. Complex statistical analysis requires a larger sample size or, alternatively, lowers power. For example, a smaller sample is required when running a *t*-test versus an ANOVA with multiple groups. Although complex designs should not be abandoned if the research necessitates it, this certainly should be part of the consideration to insure the highest probability of success with the goals of the research.

One more way to improve power or lower sample size needs is to use parametric analyses instead of nonparametric types. While the differences between the two types of analysis are marginal when sample sizes are large, there are noticeable differences when dealing with small samples. Since generally the problem is the sample size is too small, parametric analyses should be chosen if possible. Such advice does come with the caveat that small samples often do not fit the assumptions of parametric tests, so caution is necessary to insure that the attributes of the sample are examined for compliance with such assumptions.

It is important to note that even if an *a priori* power analysis comes up short of the recommended .80, that in itself is not a reason to abandon the research project. If the value is still lower than the .80 or other value selected by the researcher after every effort has been made to improve power, the research can still move forward with the research but should note the power issue as a potential significant limitation. Also, if the null-hypothesis ends up being retained, the researcher would need to explain that this could be attributed to the study being underpowered. Researchers should still feel confident in submitting such studies for publication because much can be learned from the design, implementation, sampling, analysis, and findings, or lack thereof. And since there still is a limited amount of aviation literature available to the research community, such studies can be enlightening on how to design and conduct future studies as well as identifying areas that call for additional investigation. See appendix A for a checklist on ways to improve power or reduce sample size.

The findings here can also assist individuals other than researchers. The evidence presented here should serve as an encouragement to journal editors and reviewers to pursue the recommendations of the APA Task Force on Statistical Inference by requiring the inclusion of evidence of power analysis and effect sizes in submissions. A wide range of journals now require such data in all submissions (Ferrin et al., 2007). This movement could help standardize the reporting of research making it easier for interpretation and evaluation results. This should help align aviation research with mainstream research. Perhaps the most positive effect would be that "with an understanding of effect size estimates and confidence intervals, […]

researchers can go beyond the reporting of statistical significance (*p*-value) and report on practical significance" (Ferrin et al., 2007, p. 99) thus findings within aviation studies would be able to have enhanced meaning and applicability by allowing stakeholders to go beyond the typical dichotomous findings of hypothesis testing to find deeper, more pragmatic utility of results and conclusions. Editors and reviewers could use the findings and recommendations in this study to analyze the appropriateness of methods used by researchers. Even if a study is found to be underpowered, reviewers and editors should determine if the researcher recognized this limitation and made efforts to mitigate its effects on the study. As long as any limitations are properly recognized, the article should still receive consideration for publication keeping in mind the potential utility of the study in expanding the research literature even if the study is underpowered.

In sum, the aviation research studied here appears to fall short of minimum desirable statistical power levels. This body of research infrequently discussed or calculated power and commonly neglected to present effect sizes. These facts call into question the sample size strategies used in these studies. Further, the validity of the conclusions made upon statistical analyses could therefore be debatable. In spite of this, aviation research does appear to be on par on most levels with current research in other subject areas. As these other fields call for higher standards for the reporting of research findings, aviation research must keep pace by doing the same. Moreover such improvements in research design and data analysis will provide for more complete, easier to understand, replicable, and meaningful research.

Recommendations

The findings of this research call for suggestions for consideration and for future investigation. These include:

1. An expanded study should be conducted on a wider range of aviation publications that includes subject areas such as psychology and human factors journals that likely would have large amounts of statistical analysis.

2. Editors and reviewers of aviation research journals should begin the discussion of raising data reporting standards to include appropriate sample size calculation, power analysis, the inclusion of effect sizes, and additional standards recommended by the APA Task Force on Statistical Inference.

3. Editors and reviewers of aviation research journals may want to begin to accept research methods and best practices articles. These could help disseminate research-based guidance on how to conduct power analyses, calculate effect sizes, use and interpretation of confidence intervals, and how to appropriately cite statistical findings. Such "how-to" articles are common in many other fields of study and are certainly scholarly in nature as they are entirely rooted in the available research literature.

4. Aviation researchers should include evidence based reasoning for the selection of sample size, appropriate consideration of power, and the considered effect size. Further, researchers should insure that they report their statistical findings in a recognized, standard format (e.g. APA). If a study is underpowered, this should be clearly explained as a limitation and efforts to mitigate the effects of this on the study should be discussed.

5. Considering the small sample sizes that are common in some aviation studies, there should be a call for collaboration among aviation programs to further enhance the body of research by boosting available sample sizes. These enhanced samples may provide more compelling results and perhaps make findings more generalizable.

6. Research sponsors such as the FAA and NASA should require the reporting of power and effect size for funded research projects.

7. Editors should supply a checklist of requirements to submitters that would include standards for statistical reporting, e.g. the inclusion of power and effect size as well as reporting all data in a standardized (APA) format.

8. Further research should be conducted into the quality of statistical reporting in aviation research.

References

American Psychological Association (APA). (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Aguinis, H., Beaty, J., Boik, R., & Pierce, C. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*(1), 94-107.

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology, 23*(3), 399-406.

Borkowski, S. C., Welsh, M. J., & Zhang, Q. M. (2001). An analysis of statistical power in behavioral accounting research. *Behavioral Research in Accounting, 13*, 63-84.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*(3), 145-153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Psychological Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Coladarci, T., Cobb, C., Minium, E., Clarke, R. (2007). *Fundamentals of statistical reasoning in education* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Dattalo, P. (2008). *Determining sample size: Balancing power, precision, and practicality*. New York: Oxford University Press.

Fagley, N. S. (1985). Applied Statistical Power Analysis and the Interpretation of Nonsignificant Results by Research Consumers. *Journal of Counseling Psychology, 32*, 391-396.

Ferrin, J., Bishop, M., Tansey, T., Frain, M. Swett, E., & Lane F. (2007). Conceptual and practical implications for rehabilitation research: Effect size estimates, confidence intervals, and power. *Rehabilitation Education, 21*(2), 87-100.

Fisher, R.A. (1966). *The design of experiments* (8th ed.). Hafner: Edinburgh, UK. (Original work published 1935)

Jones, A., & Sommerlund, B. (2007). A critical discussion of null hypothesis significance testing andstatistical power analysis within psychological research. *Nordic Psychology, 59(*3), 223-230.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kosciulek, J., & Szymanski, E. (1993). Statistical power analysis in rehabilitation counseling research. *Rehabilitation Counseling Bulletin, 36*(4), 212.

Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in Sociology. *Social Forces, 84*(1), 1-24.

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: how effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology, 28*(2), 151-160.

Sedlmeier,P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309-316.

Stevens, J. P. (2007). *Intermediate statistics: A modern approach* (3rd ed.). New York: Lawrence Erlbaum Associates.

Spanos, A. (1999). *Probability theory and statistical inference: econometric modeling with observational data*. Cambridge, UK: Cambridge University Press.

Spybrook, J. (2008). Are power analyses reported with adequate detail? Evidence from the first wave of group randomized trials funded by the Institute of Education Sciences. *Journal on Research Educational Effectiveness, 1*, 215-255.

University of New England. (2000). *What α-level*? Retrieved from http://www.une.edu.au/WebStat/unit_materials/c5_inferential_statistics/what_alpha_leve.html

Blank