Journal of Agricultural Education, 58(1), 300-312.
https://doi.org/10.5032/jae.2017.01300

Power of Statistical Tests Used to Address Nonresponse Error in the *Journal of Agricultural Education*

Donald M. Johnson[1] & Catherine W. Shoulders[2]

Abstract

As members of a profession committed to the dissemination of rigorous research pertaining to agricultural education, authors publishing in the Journal of Agricultural Education (JAE) must seek methods to evaluate and, when necessary, improve their research methods. The purpose of this study was to describe how authors of manuscripts published in JAE between 2006 and 2015 tested for nonresponse error. Results indicated that none of the studies' tests had acceptable power to detect small effect sizes, 14.3% had acceptable power to detect medium effect sizes, and 43% of the studies' tests did not have acceptable power to detect large effect sizes. These findings suggest that while authors frequently find no difference between respondents and others, the tests used to detect these differences are often not powerful enough to do so, leading to higher than acceptable risk for Type II error. Using the theory of planned behavior as a framework, we highlight these findings to spur change within the profession's expectations of reporting statistical power when testing for nonresponse error and offer a primer to improve researchers' perceived behavioral control over reporting power. We also offer specific suggestions for conducting and reporting the results of tests for nonresponse bias.

Keywords: Power, Nonresponse error, Statistical tests

Introduction

According to the *Journal of Agricultural Education's* (*JAE)* philosophy and policies, the journal exists to provide a vehicle for the wide dissemination of "results of research…in agricultural education" (AAAE, n.d., para. 1). The journal's policies related to article retraction suggest that research articles published within the journal are assumed to be of high rigor. This assumption is supported by the manuscript's review process, which asks reviewers to evaluate manuscripts based on rigor, including the completeness and correctness of the study's methods and procedures. The value of quality research has also been established by experienced reviewers, who stated, "Quality research contributes to the body of knowledge…Manuscripts that fail to adequately demonstrate these characteristics are not useful" (Roberts, Barrick, Dooley, Kelsey, Ravin, & Wingenbach, 2011, p. 1-2).

The American Association for Agricultural Education (AAAE), which is the parent organization for the *JAE*, has suggested the reporting of effect sizes for statistical significance in quantitative data analysis as an aspect of maintaining rigor (AAAE, 2016), as they "enable researchers to judge the practical significance of quantitative research results" (Kotrlik, Williams, & Jabor, 2011, p. 132). This requirement of AAAE research conference submissions echoes the APA's value of effect sizes, as the 2009 guidelines note, "it's almost always necessary to include some measure of effect size in the Results section" (p. 34). According to the Neyman-Person method of statistical inference, the effect size displays the degree to which an alternate hypothesis

[1] Donald M. Johnson, Department of Agricultural Education, Communications and Technology, University of Arkansas, 205 Agriculture Building, Fayetteville, AR 72701. Email: dmjohnso@uark.edu
[2] Catherine W. Shoulders, Department of Agricultural Education, Communications and Technology, University of Arkansas

(H_1), which states there is a difference in the mean of a particular characteristic between two or more groups within a population, is different from the null hypothesis (H_0), which states there is no difference in the mean of a particular characteristic between two or more groups within a population (Cohen, 1992). While effect sizes for each statistical test are reported on continuous scales that range upward from zero, with an effect size of zero being equal to the H_0, Cohen (1988) proposed operational definitions of small, medium, and large effect sizes for each scale which allow researchers to compare effect sizes between studies using different statistical tests. According to Cohen (1992), these definitions were created so that:

> Medium [effect sizes] represent an effect likely to be visible to the naked eye of the careful observer. (It has since been noted in effect size surveys that it approximates the average size of observed effects in various fields.) I set small [effect size] to be noticeably smaller than medium but not so small as to be trivial, and I set large [effect size] to be the same distance above medium as small was below it. (p. 156)

Cohen's operational definitions of effect size have been widely adopted for general use across disciplines (Cohen, 1992), including within the discipline of agricultural education (Kotrlik et al., 2011).

However, the ability of a researcher to detect a small, medium, or large effect size, which is termed statistical power, depends on the sample sizes of the groups from which the statistical test was calculated. The power of a significance test can determine whether a researcher is able to find a small, medium, or large effect size that exists within a population, or whether that researcher is more likely than not to miss that effect and instead commit Type II error. Because population sizes in agricultural education vary, it is possible that some researchers fail to obtain sufficient power to detect differences between groups. While committing Type II error leads to more conservative findings and conclusions when determining the impact of treatments that could influence future practices, Type II error can lead to particularly concerning results with regard to nonresponse error. Nonresponse error, which results from failing to include in the sample participants whose responses are representative of all members of a population (Lindner, Murphy, & Briers, 2001), may occur when less than 100% of an appropriately acquired sample responds. If nonrespondents differ from respondents on variables relevant to the survey, generalizing the findings directly from the respondents to the population results in biased estimates of population parameters. According to Reio (2007), the extent of this nonresponse bias can be calculated as:

Nonresponse bias = Proportion of Nonrespondents ($M_{\text{respondents}} - M_{\text{nonrespondents}}$)

Two important insights can be derived from the formula for nonresponse bias. First, a high response rate is not an entirely adequate protection against nonresponse bias if large differences exist between respondents and nonrespondents (Martin, 2004). Second, a low response rate may provide an unbiased estimate of the population parameters if there is little or no difference between respondents and nonrespondents (Martin, 2004). For example, in a survey with an 80% response rate (proportion of nonrespondents = .20) and a mean of 4.50 (on a 1 to 5 scale) for respondents and a mean (although unknown) of 1.50 for nonrespondents, the nonresponse bias is equal to -0.60, resulting in an actual population mean of 3.90, not the biased estimate of 4.50. Conversely, in a survey with a 20% response rate (proportion of nonrespondents = .80) and a mean of 4.50 for respondents and a mean of 4.40 (although, again, unknown) for nonrespondents, the nonresponse bias is equal to -0.08, and the sample mean for respondents is a relatively unbiased estimate of the population mean of 4.42, despite the low response rate.

Methods to address potential nonresponse bias include comparing respondents to the population on appropriate characteristics, comparing respondents to nonrespondents on specific known characteristics, comparing early to late respondents, and gaining access to nonrespondents in a manner known as "double dipping" (Miller & Smith, 1983). Roberts et al. cited nonresponse as a common threat to research within *JAE* (2011), and because of the numerous methods available to address nonresponse error, stressed that authors "account for non-response error and…select methodologies based on best practices in the social sciences, not on perceived ease" (p. 3). Lindner et al. (2001) determined that of the 114 manuscripts that attempted to control for nonresponse error published in *JAE* between 1990 and 1999, 75.4% did not find any difference between respondents and nonrespondents (which could have been represented by late respondents).

The AAAE membership has previously established the value of examining the procedures used in conducting research, with frequently cited publications recommending authors devote time to improving their knowledge and abilities within research methods (Lindner et al., 2001; Miller & Smith, 1983; Roberts et al., 2011). Cohen (1992) has identified power as a weak area within social science research and noted that while "there is no controversy among methodologists about the importance for power analysis…It is not at all clear why researchers continue to ignore power analysis" (p. 155). Because of the varied sample sizes used within agricultural education research and the frequency with which response rates total less than 100%, the ability for researchers to confidently and accurately report differences between respondents and nonrespondents with acceptable power must be established.

Theoretical Framework

This study examined the power of statistical methods used to test for nonresponse error within manuscripts published in *JAE* between 2005 and 2016 using the theory of planned behavior as a framework. The theory of planned behavior states that an individual's behavior is a result of his or her intention to perform the behavior, which is influenced by his or her attitude toward the behavior, the subjective norm regarding the behavior, and the individual's perceived control over the behavior (see Figure 1) (Ajzen, 1991).

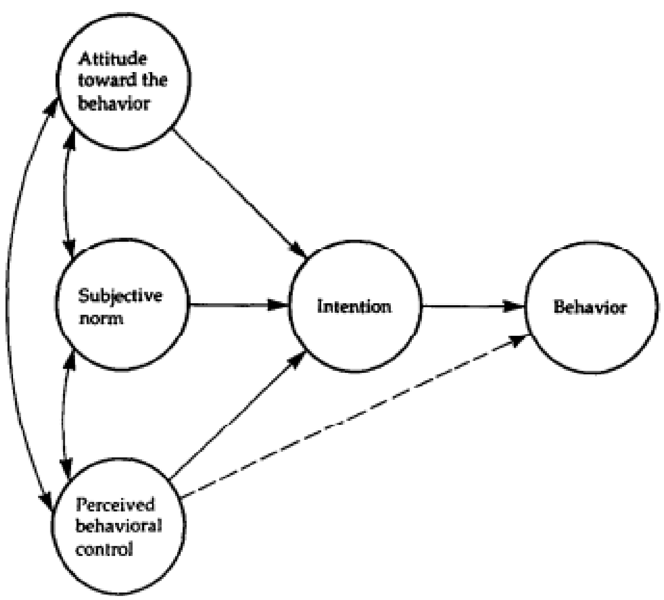

Figure 1. Theory of Planned Behavior (Ajzen, 1991).

A person's intention to perform a behavior indicates "how hard [her or she is] willing to try, how much of an effort they are planning to exert, in order to perform the behavior" (Ajzen, 1991, p. 181). Intention is shaped by an individual's perceptions regarding his or her ability to perform the behavior, namely "the ease or difficulty of performing the behavior" (Ajzen, 1991, p. 183). Individuals who are confident in their ability to perform a behavior are likely to put more effort in to performing that behavior; similarly, improving a person's perceived behavioral control and therefore their behavioral intentions regarding a particular behavior can be increased by increasing available resources or knowledge or reducing unfamiliar elements within the behavioral situation (Ajzen, 1991). Attitude toward a behavior "refers to the degree to which a person has a favorable or unfavorable evaluation or appraisal of the behavior in question (Ajzen, 1991, p. 188). This construct, along with the subjective norms regarding the behavior, which refer to the "perceived social pressure to perform or not perform the behavior" (Ajzen, 1991, p. 188) can combine with perceived behavioral control to determine an individual's intention to perform a behavior.

Within the context of this study, we use the theory of planned behavior to posit factors which would increase researchers' likelihood to report power during analysis of differences between respondents and nonrespondents. Because of the previously established value AAAE members place on quality research, we assume researchers hold positive attitudes regarding performing statistical tests with appropriate effect sizes which accurately identify differences within the population (in other words, the researchers would rather not commit Type II error if their actions could reduce its likelihood in occurring). Cohen (1992) has noted an absence of establishing power in a variety of social science disciplines, and posits that reasons for its omission may be due to "the low level of consciousness about effect size" (p. 155) or perhaps "researchers find it too complicated, or do not have at hand…reference material for power analysis" (p. 156). These potential barriers reduce researchers' perceived behavioral control in calculating and reporting power when testing for nonresponse error. And, as Cohen (1992) mentioned, because few researchers are reporting power calculations within their manuscripts, there is little expectation or pressure from the research society to do so. The theory of planned behavior suggests that by shifting the subjective norm to include an expectation of power reporting (as has occurred following the publication of previous manuscripts examining research procedures within the profession [Lindner et al., 2001]), and by improving researchers' perceived behavioral control in reporting power calculations via reducing unfamiliar aspects of the concept, researchers' intentions, and ultimately, their behaviors, regarding the reporting of power during tests for nonresponse error can be altered.

Conceptual Framework

To offer readers a framework for the concept of statistical power (Camp, 2001) and reduce any knowledge-related barriers (Ajzen, 1991) to considering power when reporting results of tests used to address nonresponse error, we offer a primer and example. When researchers attempt to determine the differences between two groups of people within a population via samples of individuals from those groups, they are attempting to extrapolate the results from those two samples to the larger population. Therefore, a group that is not representative of the population will be unable to produce results representative of the entire population, particularly if the variable of interest is linked to a characteristic within the group. We will explain this concept using an example wherein a researcher is comparing those that responded and those that did not respond to a survey inquiring about respondents' heartrates. It is possible that nonrespondents were not as healthy as respondents, and were therefore unable to summon the energy to respond to the survey. Had they responded, their responses to the survey about their heartrates would have likely been different than responses from those that were healthy enough to respond. Without comparing respondents to nonrespondents in some way, the researcher would not be able to determine whether the responses received were likely to be representative of the population. When comparing respondents to nonrespondents, the researcher certainly hopes to detect differences between the two groups, should a difference be present. Failing to determine a difference between the two groups would lead the researcher to commit Type II error and lead readers to believe the respondents were characteristic of the entire population when in fact they were not. Type II error in comparing respondents to nonrespondents leads to overgeneralization of findings to a population not represented by the sample. Cohen (1988, 1992) established that research with a 20% chance of committing Type II error is acceptable, which has been a widely adopted standard across disciplines (Field, 2009).

Differences between groups can occur at different levels of magnitude. In our example, the difference between heartrates of healthy respondents and those that did not respond because they were ailing in health would be much smaller than the difference between heartrates of healthy respondents and those that did not respond because they were dead. Further, the difference between respondents and those that did not respond because they were busy doing physical activities would likely be even smaller. The magnitude of the difference on a variable within the population is the effect size, and as Cohen (1988, 1992) has established, the effect size can be small (healthy respondents to physically active nonrespondents), medium (healthy respondents to unwell nonrespondents) and large (healthy respondents to dead nonrespondents). A statistical test's ability to detect these effect sizes is known as power, and is the opposite of Type II error. Because an acceptable Type II error risk is .20, the minimum acceptable power is 1 - .20, or .80 (Cohen, 1998; 1992; Field, 2009).

While researchers hope to have at least an 80% chance of detecting a difference present between respondents and nonrespondents and only a 20% chance of missing that difference, their ability to do so depends on the sample sizes they use to compare groups. Smaller effect sizes are less apparent than larger effect sizes, and therefore require larger sample sizes for detection. Smaller sample sizes reduce the researcher's ability to detect a difference, thus reducing power and increasing the risk of Type II error. Power also depends on the statistical test used, as effect size calculations differ by test, and on the significance criterion (α [alpha level]). While the standard alpha level for social science research is .05, .10 can be used in "circumstances in which a less rigorous standard for rejection is desired" (Cohen, 1992, p. 156). Cohen has established appropriate sample sizes required in order to have an 80% chance of detecting effect sizes of each magnitude and at each alpha level and for each statistical test. Table 1 displays Cohen's recommended sample sizes per group for tests comparing two means, as two-tailed *t*-tests are most commonly used when comparing respondents to nonrespondents or early respondents to late respondents (see Table 1).

Table 1

N per group for small, medium, and large effect sizes for two-tailed t-tests at power = .80 for α = .01, .05, and .10 (Cohen, 1992, p. 158)

α								
.01			.05			.10		
Sm	Med	Lg	Sm	Med	Lg	Sm	Med	Lg
586	95	38	393	64	26	310	50	20

To illustrate the importance of sample size in the power of tests of significance, we will refer back to our example. To determine the heartrate differences between respondents and nonrespondents, a researcher may use a two-tailed independent samples *t*-test. If, among the population, the researcher anticipates a medium effect size (in our example, a medium effect size was representative of nonrespondents being those in poor health), he or she must have data from 64 respondents and 64 nonrespondents in order to have an 80% chance of finding a difference at the .05 level. If, with 64 cases per group, *t* is not significant, either the actual population effect size is smaller than .50 (a medium effect size for *t*-tests, [Cohen, 1992]), or a Type II error has been committed (of which there is a 20% chance). Illustrating the extreme, we will point out the case if the researcher had fewer than 26 cases per group, at which point he or she would have a less than 80% chance of even detecting the difference in heartrate between the living and deceased (a large effect size, $\geq$ .80).

The risk for committing a Type II error in tests of nonresponse error within agricultural education research may be higher than the accepted 20%, as sample sizes used to address nonresponse error can be as small as 10% of the responding sample (Miller & Smith, 1983). Lindner, et al. (2001) found that response rates for studies within *JAE* between 1990 and 1999 ranged from 28% to 100%, with the average being 81.6%. Nearly 70% of studies had a response rate of less than 100%, establishing the need to test for nonresponse error (Lindner et al., 2001). In the Lindner et al. (2001) study, nonresponse error was addressed by comparing early to late respondents in 31.3% of the studies, by double dipping with a sample of nonrespondents in 18.7% of the studies, by comparing respondents and nonrespondents on characteristics known *a priori* in 2.3% of the studies, and by comparison between respondents and the population on characteristics known *a priori* in 0.9% of the studies. Nearly 47% of the studies with less than a 100% response rate did not attempt to test for nonresponse error.

Purpose and Objectives

The purpose of this study was to describe how authors of manuscripts published in *JAE* between 2006 and 2015 tested for nonresponse error. To achieve this purpose, we developed the following objectives:

1. to describe methods used to test for nonresponse error in manuscripts published in *JAE;*
2. to describe references cited in support of authors' method of testing for nonresponse bias;
3. to describe the frequency with which manuscripts included data and methodological details required to allow readers to conduct *post hoc* power analyses of the authors' statistical tests for testing for nonresponse bias; and
4. to determine the power of tests used in testing for nonresponse error within manuscripts published in *JAE*.

Methods

Articles were identified by manually examining all 526 articles published in the *JAE* between 2006 and 2015. This descriptive study examined 127 articles out of the 526 total articles published in the *JAE* between 2006 and 2015, inclusive, where authors reported an attempt to test for nonresponse error. In four articles, the researchers reported unsuccessful attempts to follow-up with nonrespondents and cautioned readers not to generalize beyond the respondents; these articles were included only in the analyses of the methods used to test for nonresponse bias and references cited (objectives 1 and 2). Four articles reported surveys conducted with two populations where efforts were made to test for nonresponse error with each population. For objectives 3 and 4, each test for nonresponse bias was considered as a unit of analysis ($N = 127$).

A standardized coding sheet was developed and used to guide data collection from each article, with the following information collected: (a) article identification number, (b) response rate(s), (c) method(s) of testing for nonresponse bias (as described by Lindner et al., 2001), (d) reference(s) cited in support of method(s) used, (e) variable(s) used in statistical comparison(s), (f) statistical test(s) used, (g) test alpha level, (h) expected effect size, (i) sample size for each group, (j) obtained test statistic or *p*-value, (k) whether or not a statistically significant difference was found, and (l) whether statistical power was reported. The title page and the methods section of each identified manuscript were printed and keyed to the identification number on the coding sheet to allow verification of all data. All article coding was completed by one researcher. The second researcher used the same coding sheet to examine a random sample of 13 (10.2%) articles and a 100% agreement was achieved for all items coded, providing evidence of the consistency and reliability of the data coding process.

In order for a reader to calculate statistical power *post hoc*, the author must report specific statistical data in the article: (a) the statistical test(s) used, (b) the number of subjects in each comparison group, (c) the alpha level of the statistical test(s), and (d) the effect size of the difference anticipated in the population (Faul, Erdfelder, Lang, & Buchner, 2007). For this study, in cases where the authors reported the statistical tests(s) used and the number of subjects in each comparison group but did not report the alpha level, the customary alpha level of .05 was assumed. Additionally, in cases where the authors did not report the anticipated effect size, the standard small, medium, and large values offered by Cohen (1992) were assumed in order to allow for any possible researcher effect size expectations to be examined. This resulted in 35 (27.6%) tests of nonresponse bias where statistical power could be calculated *post hoc*. G*Power Version 3.1.9.2 (Faul et al., 2007) software was used to calculate statistical power and results were confirmed by comparison with Cohen's (1988) power tables.

Findings

In order to ensure an investigation into the study's objectives was warranted, we first examined the response rates of studies reported in JAE manuscripts between 2006 and 2015. The mean response rate for the 131 instances where researchers attempted to test for nonresponse error was 56.3% (SD = 18.0); the median response rate was 57.0% with an interquartile range of 28.0% (Q1 = 43.3% and Q3 = 71.3%).

Objective 1 sought to describe methods used to test for nonresponse error in manuscripts published in *JAE* (see Table 2). The majority compared early to late respondents (74%, $n = 97$), while double-dipping with nonrespondents was the second-most frequent method. Comparing respondents to a known population, regressing days to respond, and using multiple methods were each employed to a lesser degree.

The use of these methods was supported with literature from within and outside of the discipline (see Table 3). Lindner et al.'s 2001 manuscript, which was published in *JAE* and proposed three protocols for handling nonresponse issues, was cited most frequently. A manuscript by Miller and Smith (1983), which outlines nonresponse methods and was cited by Lindner et al., was cited in approximately one-third of the manuscripts. Various editions of *Introduction to Research in Education*, by Ary, et al. (1996, 2014) were also cited with some frequency. Almost 10% of articles did not cite any reference in support of their methods for testing nonresponse bias.

Table 2

Methods used to Test for Nonresponse Bias in Articles Published in the Journal of Agricultural Education, 2006 - 2015

Method	*f*	%
Compared early to late respondents	97	74.0
Followed up with a sample of nonrespondents	20[a]	15.3
Compared respondents to population on characteristics known *a priori*	8	6.1
Regressed days-to-respond on selected variables	3	2.3
Compared early to late respondents AND followed up with a sample of nonrespondents	4	3.0

Note. Based on 131 tests reported in 127 articles. [a]Includes four (4) cases where attempts to follow up with nonrespondents were unsuccessful.

Table 3

References Cited to Support Methods used to Test for Nonresponse Bias in Articles Published in the Journal of Agricultural Education, 2006 - 2015

Reference Cited	*f*	%[a]
Lindner, Murphey, & Briers, 2001	76	58.0
Miller & Smith, 1983	45	34.4
Ary, et al. (various editions and co-authors)	13	9.9
Other	10	7.6
None	13	9.9

[a] Percentages total more than 100% because some authors cited multiple source

Authors included in their manuscripts various degrees of detail regarding how those methods of testing for nonresponse bias were carried out (see Table 4). Less than half of the manuscripts detailed the specific variables on which groups were compared. Slightly more than half of the manuscripts included which statistical test was used in comparing groups, while slightly less than half listed the number of respondents within each comparison group. Less than 20% of manuscripts reported the alpha level of the statistical test or the obtained test statistic or its associated probability. None of the articles stated the statistical power of the test or the expected effect size. Nearly 92% (n = 113) of the statistical tests found no difference between groups when testing for nonresponse bias.

Table 4

Details of Analysis in Tests for Nonresponse Bias Reported in Articles Published in the Journal of Agricultural Education, 2006 - 2015

	Reported in Article	
Detail of Analysis	f	%[a]
1. Specific variable(s) used for comparison	57	44.9
2. Statistical test used in comparison	66	52.0
3. Alpha level of test(s)[b]	25	19.7
4. Number of respondents in comparison group 1	61	48.0
5. Number of respondents in comparison group 2	59	48.5
6. Expected effect size	0	0.0
7. Test statistic or associated probability	22	17.3
8. Statistical power of test	0[c]	0.0

[a] Percentages are based on 127 tests for nonresponse bias reported in 123 articles. [b]Where alpha levels and effect sizes were not reported, we assumed the customary alpha level of .05 and analyzed data for effect sizes of small, medium, and large values (Cohen, 1992) in order to allow for any possible researcher expectations to be examined. [c]In four analyses, authors reported a concern about the statistical power of their test, but did not provide the actual power coefficient.

None of the articles contained all the information necessary for a reader to calculate the statistical power of tests of nonresponse bias. However, again, by assuming an alpha level of .05 and using the small, medium, and large effect sizes specified by Cohen (1988), *post hoc* power analysis could be conducted for 35 (27.6%) tests. None of these tests achieved the acceptable power of .80 at the small effect size, five (14.3%) achieved power of .80 at the medium effect size, and 20 (57%) achieved power of .80 at the large effect size. The average likelihood these tests would detect a small effect size was 15% (see Table 5). Manuscripts had a mean power of .57 at the medium effect size, meaning that, on average, the odds of a manuscript accurately detecting a medium effect size was slightly greater than 50/50. The average manuscript power at the large effect size was .86, which exceeds the minimum acceptable threshold of .80.

Table 5

*Post Hoc Analyses of the Statistical Power for Tests of Nonresponse Bias (*n *= 35) Reported in the Journal of Agricultural Education, 2006 - 2015*

	Statistical Power			
Effect Size[a]	*M*	*SD*	Median	IQ_{range}
Small (*d* = 0.20)	.15	.07	.13	.05
Medium (*d* = 0.50)	.57	.21	.56	.23
Large (*d* = 0.80)	.86	.15	.92	.15

[a] Descriptors based on Cohen (1988).

Conclusions/Implications/Recommendations

Fifteen years after Lindner et al. (2001) published results that stated 46% of *JAE* manuscripts did not attempt to test for nonresponse error and provided recommendations for doing so, the state of professional practice in agricultural education research has improved to the point where controlling for nonresponse error is an essential component of manuscript quality (Roberts et al., 2011). As Ajzen (1991) posited, by altering the social norms to include an expectation of testing for nonresponse error and offering familiar methods for doing so, researchers began including tests for nonresponse error in their manuscript writing behaviors more frequently. However, as Cohen (1992) lamented with regard to the work of social scientists in general, authors publishing within *JAE* omitted the subject of power from their statistical calculations. None of the articles published between 2006 and 2015 stated the statistical power of their tests, although 92% of them stated no difference was found (a statement that can only be made with 80% accuracy if the power ≥ .80). Less than one-third of the manuscripts included sufficient detail for the reader to calculate the power of the test used even when assuming the test alpha level and using standard effect sizes.

When power was calculated and compared to the minimum acceptable threshold of .80 suggested by Cohen (1988), we found that none of the statistical tests of nonresponse bias were adequate to detect small effect sizes, which were those that are smaller than medium, but not trivial (Cohen, 1992). Eighty-five percent of the tests were not adequate to detect medium effect sizes, which are those that are observable to the trained eye (Cohen, 1992); in our previously used example, medium effect size was likened to that of the heartrates between healthy individuals and physically unwell individuals. Close to half of the tests were not adequate to detect even large effect sizes (such as the difference in heartrate between alive and deceased individuals), leading to a very likely chance that the lack of differences found between respondents and others could have actually been the result of Type II error. The dissemination of these findings offer a catalyst to alter the profession's subjective norm regarding the reporting of power when testing for nonresponse error (Ajzen, 1991), as researchers seek to fulfill the expectations of their peers by minimizing their risk for committing Type II error. In order to increase researchers' perceived behavioral control in calculating and reporting power, and in reducing Type II error, we propose several recommended practices when testing for nonresponse error (Ajzen, 1991).

When conducting tests for nonresponse bias, researchers should prioritize statistical power over the employment of previously used methods as they determine appropriate sample sizes for groups of nonrespondents or late respondents. While Miller and Smith (1983) recommended sampling 10-20% of nonrespondents, indiscriminate use of this method without consideration of the number of nonrespondents provides no guarantee with regard to power, and therefore with regard to protecting against Type II error.

We also recommend that researchers determine and report the power of their test(s) of nonresponse bias in order to inform the reader of the risk of Type II error. Researchers should include necessary components for the reader to calculate power on his or her own, including the statistical test used, the number of respondents in each group, the alpha level used, and the anticipated effect size. While no published rule exists on the methods for anticipating effect sizes within a population, we recommend researchers cite justification from the existing literature base for their estimates whenever available. The following may serve as an example for reporting the results of tests for nonresponse bias:

To test for nonresponse bias, early respondents (those responding prior to the third mailing, $n = 102$) were compared to late respondents ($n = 59$), on the variable 'commitment to teaching as a career', using a two-tailed independent *t*-test at the .10 alpha level; the power of the test was .92 for a medium effect (*Cohen's d* = 0.50 [Cohen, 1988]). There was no significant difference between early ($M = 4.43$, $SD = 0.82$) and late ($M = 4.34$, $SD = 0.87$) respondents, $t(157) = 0.65$; $p = .52$. Thus, the findings were generalized to the population (Miller & Smith, 1983; Linder et al., 2001).

We believe researchers should be more open to the possibility that respondents and nonrespondents are different in meaningful ways and that nonresponse bias is a threat to generalizability until we have sufficient evidence to the contrary. Subjects who do not respond to a survey (or respond late) already differ from those who do respond (or respond early) in at least one dimension. Despite this difference, we traditionally assume respondents and nonrespondents are not different and then require overwhelming evidence ($p < .05$) of this difference before we are willing to acknowledge the possibility of nonresponse bias. Researchers are accustomed to assuming no difference exists between groups until proven otherwise (H_0) when hypothesis testing in order to yield conservative results and avoid mistakenly generalizing results beyond the sample. However, this same assumption, when testing nonresponse bias, actually yields less conservative results (finding no difference where there is one yields to overgeneralization of the study's results). Therefore, we recommend researchers not reflexively use the .05 alpha level in testing for nonresponse bias. When selecting an alpha level, researchers directly control the probability of committing a Type I error and indirectly the probability of committing a Type II error (Mitchell & Jolley, 2010). In introductory research methods courses, many of us were taught that, instead of blindly following convention, we should base our alpha level on the relative consequences of committing a Type I versus a Type II error. In testing for nonresponse bias, a Type I error incorrectly leads researchers to limit their findings, conclusion, and recommendations to the respondents. Conversely, a Type II error incorrectly leads researchers to generalize from the respondents to the population when this generalization is not appropriate. We argue that, given the state of the science in agricultural education research, Type I errors are generally less consequential than Type II errors in this specific instance and recommend use of less conservative alpha levels (.10 or .15) in tests of nonresponse bias (Cohen, 1988).

Finally, when tests with acceptable power yield statistically significant differences between groups or when acceptable power is not attainable, researchers should caution readers against generalizing results beyond the sample of respondents to prevent overgeneralizating the findings. Studies yielding valid results of interest to the profession from a specific groups of respondents, regardless of their generalizability, can add to the body of knowledge and assist researchers as they design and conduct research. Stated differently, researchers within agricultural education should accept the proposition that meaningful results appropriately limited to only the respondents are more valuable than results inappropriately generalized to the population.

The results of this study confirm Cohen's (1992) concerns about omitting consideration of power when employing statistical tests. While the recommendations made align with those made by Cohen (1988, 1992), change in the societal norms regarding testing for nonresponse error and the inclusion of power calculations within *JAE* is in the hands of the profession's members. As Lindner, et al. (2001) stated, "the profession will verify or refute the utility of the methods proposed here" (p. 52). Regardless of the behavioral change produced, any examination into the utility of these recommendations will continue to further the profession's commitment to improvement in research methods (Lindner et al., 2001; Miller & Smith, 1983; Roberts at al., 2011).

References

AAAE. (n.d.) JAE philosophy. *The American Association for Agricultural Education.* Retrieved from http://jae-online.org/index.php/philosophy.

AAAE. (2016). AAAE Manuscript Submission Guide. Retrieved from http://aaaeonline.org.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.

Ary, D., Jacobs, L., & Razavieh, A. (1996). *Introduction to research in education* (5th ed.). Fort Worth, TX: Holt, Rinehard, & Winston, Inc.

Ary, D., Jacobs, L.C., Sorenson, C. K., & Walker, D. (2014). *Introduction to research in education* (9th ed.). Belmont, CA: Cengage Learning.

Camp, W. G. (2001). Formulating and evaluating theoretical frameworks for career and technical education. *Career and technical education research, 26*(1), 1-5.

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

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Field, A. (2009). *Discovering statistics using SPSS.* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Kotrlik, J. W., Williams, H. A., & Jabor, M. K. (2011). Reporting and interpreting effect size in quantitative agricultural education research. *Journal of Agricultural Education, 52*(1), 132-142. doi: 10.5032/jae.2011.01132

Lindner, J. R., Murphy, T. H., & Briers, G. E. (2001). Handling nonresponse in social science research. *Journal of Agricultural Education, 42*(4), 43-53. Doi: 10.5032/jae.2001.04043

Martin, E. (2004). Presidential address: Unfinished business. *Public Opinion Quarterly, 68*(3), 439-450. doi: 10.1093/poq/nfh020

Miller, L. E., & Smith, K. L. (1983). Handling nonresponse issues. *Journal of Extension*, *21*(5), 45-50.

Mitchell, M. L., & Jolley, J. M. (2010). Research design explained (7th ed.). Independance, KY: Wadsworth Centage Learning.

Reio, Jr., T. G. (2007). Survey nonresponse bias in social science research. *New Horizons in Adult Education and Human Resource Development, 21*(1/2), 48-51.

Roberts, T. G., Barrick, R. K., Dooley, K. E., Kelsey, K. D., Raven, M. R., & Wingenbach, G. J. (2011). Enhancing the quality of manuscripts submitted to the *Journal of Agricultural Education*: Perceptions of experienced reviewers. *Journal of Agricultural Education, 52*(3), 1-5. doi: 10.5032/jae.2011.03001