Journal of Agricultural Education, 60(3), 291-303
https://doi.org/10.5032/jae.2019.03291

Beyond Magic Words and Symbols: Rethinking Common Practices in Quantitative Research

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

Abstract

One of the keystones of scientific disciplines is the production and dissemination of quality, valid research. Within the Journal of Agricultural Education, traditions passed down from mentor to mentee have led to the establishment of magic words and symbols that, while used to meet criteria for acceptance for publication, actually obscure and obfuscate the research process. When used inappropriately, these terms can suggest appropriate methodology to the uninformed reader or reviewer, perpetuating the publication and dissemination of invalid research. Using the theory of planned behavior as a framework, this research note seeks to highlight some of the more common magic words and symbols used in manuscripts within the Journal of Agricultural Education and offers information to enable researchers to cease inappropriate use of these terms in an effort to enhance the validity of research published within the journal. This research note was an invited presentation at the 2019 American Association for Agricultural Education Conference and was approved for publication by the Chair of the Journal of Agricultural Education's Editing Managing Board.

***Keywords*:** validity, reliability, nonresponse error, sample selection, inferential statistics

Introduction

According to Kerlinger (1986), research is the "systematic, controlled, empirical, and critical investigation of natural phenomena" (p. 10) conducted in such a manner that "investigators can have critical confidence in research outcomes" (p.10). Kerlinger further stated that, metaphorically,

> Every scientist writing a research report has other scientists reading what he [sic] writes while he [sic] writes it. Though it is easy to err, to exaggerate, to overgeneralize when writing up one's work, it is not easy to escape the feeling of scientific eyes constantly peering over one's shoulders. (p. 11)

This statement implies that research reports, including conference papers and journal articles, must include the essential information that enables disciplinary peers and the larger scientific community to 'peer over one's shoulders' and judge the scientific quality of the researcher's methods and results, and their interpretation of these results. The use of 'magic' words and symbols in reporting quantitative research obscures others' view of these essential elements and makes critical evaluation difficult, if not impossible.

According to the Merriam-Webster online dictionary, the word "*magic* goes back to the 1300s, and originally referred to rituals, incantations, or actions thought to have supernatural power over the natural world" (n.d., para. 1). Magic words and symbols in the context of this paper are those words and symbols, commonly found in research reports and presentations, which, on the surface, seem to

[1] Donald M. Johnson is a Professor in the Department of Agricultural Education, Communications and Technology at the University of Arkansas, E111 AFLS Building, Fayetteville, AR 72701. dmjohnso@uark.edu. 479-575-2035

[2] Catherine W. Shoulders is an Associate Professor in the Department of Agricultural Education, Communications and Technology, E111A AFLS Building, Fayetteville, AR 72701. cshoulde@uark.edu. 479-575-3799

explain and clarify, but on closer examination, tend to obscure and obfuscate. However, these words continue to hold power over unsuspecting authors, reviewers, and readers.

We likely use these magic words and symbols because we learn to write research manuscripts by mimicking experienced scholars, who, in turn were once beginning scholars mimicking previous generations of experienced scholars. Thus, magic words and symbols, once in the lexicon, are difficult to dislodge. Furthermore, their widespread use builds an expectation on the part of reviewers that they *should* be included in manuscripts, ensuring that even seasoned researchers will continue to include them in the often cynical, but unfortunately not baseless, realization that their use increases the potential for a successful manuscript review.

The American Association for Agricultural Education's 2016-2020 Research Agenda stated, "members of [AAAE] have a long history of conducting high quality applied research" (p. 7). This research note seeks to highlight some of the more common "magic" words and symbols used within our research in an effort to assist reviewers in discerning between appropriate and inappropriate use of terms and methods and encourage researchers to use and report these research methods in appropriate ways.

Theoretical Framework

This article discusses common errors in agricultural education research through the lens of Ajzen's (1991) Theory of Planned Behavior (TPB). The TPB posits that reasoned actions are the result of an individual's intent to perform the behavior, which, in turn, is determined by the individual's attitude toward the behavior, the peer network's subjective norm regarding the behavior, and the individual's perceptions of his or her control over the behavior (Ajzen, 1991). In the context of this article, the behavioral change targeted is the increased use and reporting of appropriate research methods (or the discontinuation of use and reporting of inappropriate research methods) by shifting individual and peer network attitudes and by increasing perceived researcher and reviewer control in using and requiring these appropriate research methods (Figure 1). The basic premise is that when we know better, we can do better.

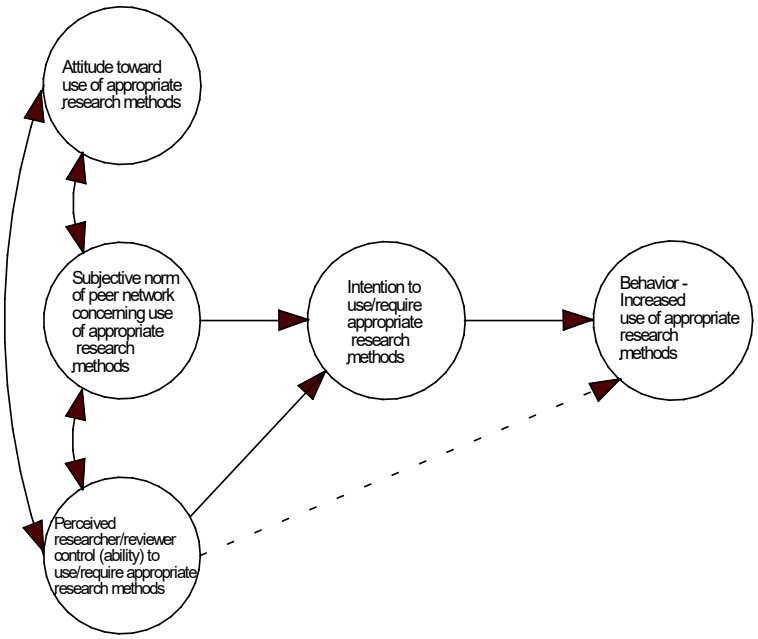

Figure 1. Theory of Planned Behavior (Ajzen, 1991) as applied to improving research methods.

Purpose

The purpose of this research note is to focus attention on some of our most commonly used magic words and symbols related to instrument validity and reliability, sample size, nonresponse error, and the use and reporting of inferential statistics. Ideally, a critical examination of these magic words and symbols will help us exorcize them from our lexicon, improving the "critical confidence" (Kerlinger, 1986, p. 10) of our research results.

Discussion

In this section of the paper, we use our combined 40-plus years of experience in writing and reviewing agricultural education research manuscripts to discuss five areas where magic words and symbols are often used by members of the profession. Specifically, these areas include our writing about instrument validity and reliability, sample selection, nonresponse error, and the use and reporting of inferential statistics.

Instrument Validity

According to Ary, Jacobs, Sorensen, and Walker (2014), "Validity is the most important consideration in developing and evaluating measuring instruments" (p. 225). Instrument validity refers to the accuracy and appropriateness of the decisions or inferences that can be made using data collected using the instrument (Ary et al., 2014; Huck, 2008; McMillan & Schumacher, 2010). Thus, instrument validity is not simply a characteristic of the instrument; rather, validity is a characteristic of the instrument, the subjects, and the conditions of administration (Gates, Johnson, & Shoulders, 2018).

The most common magic phrase related to validity is some variation of the statement, "the instrument was examined by a panel of experts and judged to be valid" (Gates et al., 2018, p. 193).

Although on the surface this phrase seems meaningful, it obscures more than it reveals. For example, this ever-popular magic phrase leaves the following questions unanswered:

1. How many "experts" were on the panel, how were they selected, and what was the nature of their expertise?
2. What specific instructions were given to the experts and what specific criteria did they judge in order to assess validity?
3. What specific type(s) of validity was (were) assessed?
4. Was the panel informed about the characteristics of the research subjects, the conditions of instrument administration, and the specific use to be made of the data generated?

Unfortunately, use of the magic phrase, "the instrument was reviewed by a panel of experts and judged to be valid," often precludes further consideration of these questions, either by the researchers themselves or by reviewers and readers. Thus, the question arises, how can the community of scholars peer over the [researcher's] shoulders and have critical confidence in this most important consideration in evaluating the quality of the study and its conclusions and recommendations in the absence of such important details?

Less common, but still troubling, is some variation of the magic phrase, "the instrument is a widely used measure of [insert construct here] (insert multiple citations here), and was, therefore, deemed valid for use in the current study" (as was found in a disconcerting number of manuscripts in the *Journal of Agricultural Education* in a study by Gates et al., 2018). In addition to the failure of this magic phrase to address any of the relevant validity questions, common use in the literature is *not* a measure of instrument validity. As an example, the Myers-Briggs Type Indicator (MBTI) is administered to over two million people each year (Stein & Swan, 2019) and a Google Scholar search for 'MBTI' (limited from 2015 to present) yielded over 7,500 hits; however, the National Research Council (1990) reported a lack of evidence firmly establishing the validity of the MTBI and reported its wide use as "curious" (p. 99). Again, widespread use does not equal validity; study specific evidence is necessary to assure the community of scholars that they can have critical confidence the instrument yields the type of data that allows the researcher to make accurate and appropriate decisions concerning the individuals completing the instrument.

According to Huck (2008), the three primary types of validity are content, criterion-related, and construct. While it is beyond the scope of this paper to present a full discussion of instrument validity, Figure 2 may prove a helpful starting point for assessing content, construct, and criterion validity.

Table 1

Key questions and evidence for three types of validity (Kerlinger, 1986; Huck 2008)

Validity Type	Key Question	Evidence
Content	Does this instrument (or scale) accurately measure a representative sample of the intended content?	Expert judgement -- Content experts -- Universe of content must be clearly specified
Construct	To what extent does the instrument (or scale) accurately measure the underlying personality or psychological construct?	Correlation with variables known to be related to construct Known group discrimination

		Confirmatory factor analysis
Criterion	Does this instrument (or scale) accurately predict scores on (or occurrence of) target criterion?	Accuracy of prediction of criterion based on instrument (or scale) score

Instrument Reliability

Borg and Gall (1983) defined instrument reliability as "the level of internal consistency or stability of [a] measuring device over time" (p. 281). Coefficient alpha is the proper procedure for assessing the internal consistency of responses to a summated scale, while the coefficient of stability is the proper procedure for assessing the stability of responses to an instrument or scale (Warmbrod, 2014).

Internal consistency, and thus coefficient alpha, is only meaningful when applied to summated scales in which responses to two or more individual items are combined to create a single overall measure of some underlying construct. When applied correctly, coefficient alpha provides an estimate of the degree to which items in a summated scale measure a single, unidimensional latent construct (Pedhazur & Schmelkin, 1991). In such instances, coefficient alpha should be calculated and reported on all appropriate summated scales and/or subscales. Warmbrod (2014) described correct procedures for reporting and interpreting scores from Likert-type scales.

When coefficient alpha is reported as a measure of instrument (or scale) reliability and the results are reported on an item-by-item basis (not as a summated scale), coefficient alpha is being used as a magic phrase and is worse than meaningless, as it leads the uninformed reader to assess the instrument's reliability via an inappropriate tool. If each item is presented as measuring a distinct variable (we will offer the example of an item assessing one's interest in eating hamburgers and another assessing one's political affiliation), the concept of internal consistency is meaningless because there is simply no reason these two items should necessarily be answered in a consistent pattern; people can enjoy hamburgers (or not) regardless of political affiliation. In such cases, what is a reader to make of any reported value for coefficient alpha? Yet, in a review of quantitative articles published in the *Journal of Agricultural Education* between 1995 and 2012, Warmbrod (2014) found that 143 (41.9%) of 344 articles incorrectly reported coefficient alpha as the measure of instrument reliability and then reported results on an individual item basis. This incorrect reporting of coefficient alpha has no place in agricultural education research; manuscripts reporting inappropriate measures of reliability are in direct violation of the profession's expectations of publishing high quality research!

The coefficient of stability (sometimes called test-retest reliability) is the correct form of reliability for instruments where results are reported on an item-by-item basis (Warmbrod, 2014). To determine the coefficient of stability, the instrument is administered to a group of individuals similar to the target sample, and then, after a delay (generally one to two weeks [Multon, 2010]), the same instrument is administered to the same individuals a second time. Item-level responses to the first and second administrations are correlated, the correlations are transformed into *z* scores, the *z* scores are averaged, and the average *z* score is transformed back to a correlation coefficient and reported as the coefficient of stability. This *z* score transformation is necessary, because correlation coefficients are ordinal level data and cannot be directly averaged. As shown in Table 1, the mean of the four respondents' test-retest correlation coefficients is $r = .50$; however, the mean *z* equivalent is 0.718, which transforms back into the equivalent mean correlation of $r = .62$.

Table 2

Results of calculating mean of correlation coefficients versus the mean of z score transformations

r	*r* to *z*	*z* to *r*
.10	0.100	
.20	0.203	
.80	1.099	
.90	1.472	
M = .50	*M* = 0.718	*r* = .62

The coefficient of stability should be calculated and reported for the entire instrument or for each applicable instrument section, as appropriate. Alternately, the range of correlations and mean correlation for the entire instrument or appropriate sections may be reported.

In establishing the coefficient of stability, it is important that the individuals used and the methods and conditions of administration be completely described and be as similar as possible to the individuals and conditions of testing used in the main study. In addition, the time between administrations should be reported.

Sample Size

Agricultural education researchers are often interested in using sample statistics to estimate population parameters. For example, a researcher might study a representative sample of agricultural science teachers to estimate the level of job satisfaction of the population of agricultural science teachers in a state. One primary concern in such situations is the size of the sample necessary to represent the population at a specified level of precision.

One widely used and cited method of determining sample size is the Krejcie and Morgan (1970) formula, or the use of tables derived from the formula. However, this formula is intended for determining samples when estimating population proportions (or percentages) at a specific probability and level of accuracy. This is clearly indicated by the actual Krejcie and Morgan (1970, p. 607) formula:

$$n = \frac{(\chi 2)(N)(P)(1-P)}{[(d^2)(N-1)+(\chi 2)(P)(1-P)]}$$

where,
n = the required sample size
χ^2 = table value of chi square for 1 *df* at desired confidence (generally 95%) level
N = the population size
P = the population proportion (generally assumed to be .50 to maximize sample size)
d = the level of accuracy of the estimate expressed as a proportion

Note quantities P and d (as well as the chi square value) in the formula refer to proportions (percentages). This is because the Krejcie and Morgan (1970) formula is intended to calculate the sample size, n, necessary to construct a confidence interval (generally $\pm$ 5%) around the sample percentage that will, in 95% (confidence level) of all samples equal to n, contain the true population percentage. Thus, use of Krejcie and Morgan to determine sample size is appropriate when the objective is to estimate population percentages from sample percentages.

In cases where the objective is to use sample statistics to estimate population means with a specified degree of confidence at a specified level of precision, the Krejcie and Morgan (1970) formula is generally too conservative (results in overly large samples); instead, Cochran's (1977) sample size formula (below) should be used.

$$n = (\frac{t\alpha\, s}{E})^2$$

Where,
n = sample size
$t\alpha$ = *t*-critical @ *alpha* (1 – *Confidence Level*) for appropriate *df*
s = estimate of population standard deviation (either from previous research or pilot study or estimated as *Number of points on scale*/6)
E = margin of error

As an example, assume we wish to estimate the population (N = 400) mean for teacher job satisfaction using a 7-point summated scale (1 = low to 7 = high). Further assume we want a 95% probability that our sample mean will be $\pm$0.25 of the true population mean. Using an estimated *SD* of 1.17 (7-point scale / 6), Cochran's sample size formula gives an initial required sample size of n_i = 85 as shown below.

$$n_i = (\frac{1.969\ x\ 1.17}{0.25})^2$$
$$= 9.21^2$$
$$= 84.8 \rightarrow 85$$

However, because this sample is drawn from a finite population (N = 400) and the initial sample size is greater than 5% of the population (0.05 x 400 = 20), Cochran (1977, p. 78) recommended adjusting the initial sample size using the following formula:

$$n_{adj} = \frac{ni}{1+\frac{ni}{N}}$$

where,
n_{adj} = the final adjusted sample size
n_i = the initial calculated sample size
N = the population size

This results in a final, adjusted sample size of n = 71, as shown below.

$$n_{adj} = \frac{85}{1+\frac{85}{400}}$$
$$= \frac{85}{1.21}$$
$$= 70.25 \rightarrow 71$$

By comparison, application of the Krejcie and Morgan (1970) sample size formula (or associated table) would result in a required sample size (n) of 196. Thus, calculation of the appropriate

sample size (n = 71) would both reduce the resources devoted to data collection and allow concentrated efforts to increase the response rate, all while achieving the specified level of accuracy.

When multiple items are included in a survey, Cochran (1977) recommended calculating required sample sizes for each item and then selecting the largest calculated sample size. Cochran also recommended increasing sample size based on the anticipated response rate so that the requisite number of responses are received.

The default sample size calculators for both Qualtrics (2018) and SurveyMonkey (n.d.) are based on the Krejcie and Morgan (1970) formula. Thus, researchers using these online tools must exercise independent scholarly judgement about the appropriateness of these calculators for their specific research project.

In addition to correct use of the Krejcie and Morgan (1970) and Cochran (1977) sample size formulas (and internet-based interactive tools based on these formulas), agricultural educators should also consider determining sample sizes based on statistical power when designing studies where inferential statistics will be used. G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) is a free, versatile, and widely used program for determining sample sizes based on statistical power. To use the program, one simply selects the statistical test, the anticipated effect size, the test alpha level, and the minimum desired statistical power (generally $\geq$ .80); once these values are set, G*Power calculates the required sample size. The calculated results are consistent with Cohen's (1988) power tables (Johnson & Shoulders, 2017). As shown in Figure 2, a three group, one-way ANOVA, tested at the .05 *alpha* level, requires an n of 159 (53 per group) to detect a medium effect (Cohen's f = 0.25) at the recommended minimum power of .80.

Of course, regardless of the sample size calculation method used, it is recommended practice to oversample base on the anticipated response rate (Israel, 2003).

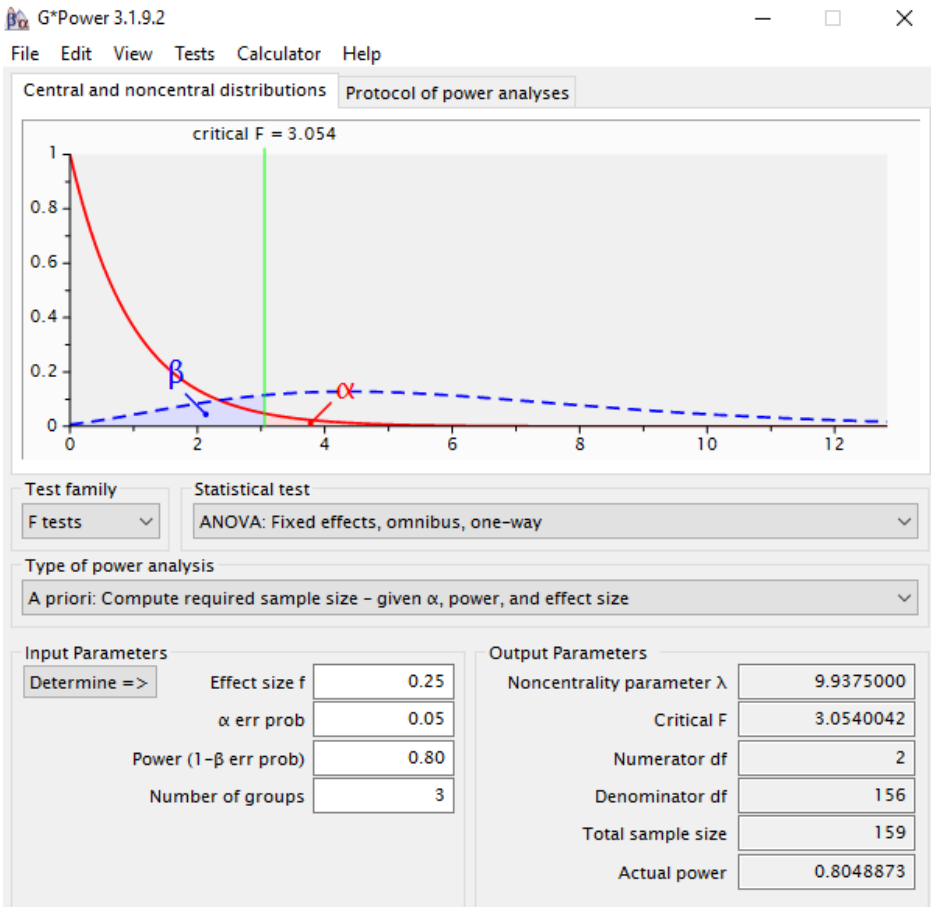

Figure 2. A priori sample size selection for a three-group, one-way ANOVA using G*Power (Faul et al., 2007).

Nonresponse Error

The potential for nonresponse error exists whenever a subset of the sampled population does not respond to a survey or other data collection effort. Actual nonresponse error exists when this nonresponding subset differs substantially from the respondents on the variables(s) of interest in the study (Reio, 2007). In such cases, nonresponse bias (NRB) can be operationalized (Reio, 2007) as,

$$NRB = P_{NR}\,(M_{Res} - M_{Nonrsp})$$

where,

P_{NR} = proportion of non-respondents

M_{Res} = the mean (or other statistic) for the respondents

M_{Nonrsp} = the mean (or other statistic) for the non-respondents

The obvious limitation to the formula is that the mean (or other statistic) for the non-respondents is unknown! However, the formula is helpful in pointing out two facts; a high response rate is not sufficient protection against nonresponse bias when there is a large difference between respondents and nonrespondents, and a low response rate may produce an accurate estimate of population parameters when there is little or no difference (Johnson & Shoulders, 2107; Rogelberg & Stanton, 2007).

In agricultural education research, the most frequently used method of testing for nonresponse bias is to compare early and late respondents (Johnson & Shoulders, 2017). In this method, respondents who complete the survey instrument after some pre-determined cut-off point are classified as late respondents and statistically compared to so-called early respondents. When no statistically significant differences are found between the two groups (which is usually the case), agricultural education researchers invoke the magic words, "Early respondents were compared to late respondents and no statistically significant differences were found; thus, the results were generalizable to the population (Linder, Murphy, & Briers, 2001)." The very use of these magic words serves to foreclose any further possible consideration by the researcher, reviewer, or reader that the 40% who did not respond are different in some important way from the 60% who did respond.

One obvious flaw in comparing early to late respondents to test for nonresponse error is that late respondents ARE respondents. According to Rogelberg and Stanton (2007),

> If late respondents differ from nonrespondents, it most likely suggests that some level of [nonresponse] bias exists. However, given that late respondents are not 'pure' nonrespondents in that they did complete the survey, being similar to respondents does not conclusively indicate an absence of [nonresponse] bias. (p. 200)

Thus, comparison of early and late respondents can only indicate the presence of nonresponse bias; it cannot prove its absence.

Johnson and Shoulders (2017) examined the statistical power of tests of nonresponse bias reported in articles published in the *Journal of Agricultural Education* between 2006 and 2015. They found that, due to small group sizes, none (0.0%) of the tests achieved an acceptable statistical power of .80 at the small effect size, 14.3% achieved a minimum power of .80 at the medium effect size, and 57% achieved a power of .80 at the large effect size. Johnson and Shoulders (2018) posited that an unknown (and unknowable) number of studies may have concluded there was no significant difference between early and late respondents (and mistakenly generalized the results) simply because they lacked the statistical power to detect small, medium and even large effect differences. The authors presented numerous recommendations to improve testing of nonresponse bias in agricultural education research.

Rogelberg and Stanton (2007) classify comparison of early and late respondents as a "lower quality" (p. 199) method of testing for nonresponse bias. Although beyond the scope of this manuscript, Rogelberg and Stanton (2007) suggested several additional methods of testing for nonresponse bias including passive nonresponse analysis, interest-level analysis, active nonresponse analysis, and worst-case resistance. Agricultural education researchers should become familiar with these methods and incorporate them as warranted.

Reporting of Inferential Statistics and Probabilities

Agricultural education researchers often seem addicted to inferential statistics and their associated *p* values! So much so that we often force the square pegs of our data into the round holes of inferential statistics. We excitedly wave our $p < .05$ or $p < .01$ or (oh my!) $p < .001$ statistical results around as if they were bumper stickers from exclusive Ivy League universities to which our sons or daughters had just been admitted. Mind you, there is nothing wrong with inferential statistics and *p* values (or even Ivy League colleges for that matter), but they should be used appropriately rather than as magic symbols.

Inferential statistics are correctly used to estimate population parameters from sample statistics. In order to make legitimate and meaningful use of inferential statistics, data must be collected from a representative sample of the population. When this is the case, researchers use inferential statistics to

make inferences from the sample back to the population with a known (or knowable) probability of being wrong (Spatz, 2018). In Figure 4, statistics (and $S_{\square}$) from a representative sample (n = 128) were appropriately used to estimate the confidence interval (± 2.5) around the population mean (μ) at the 95% confidence interval.

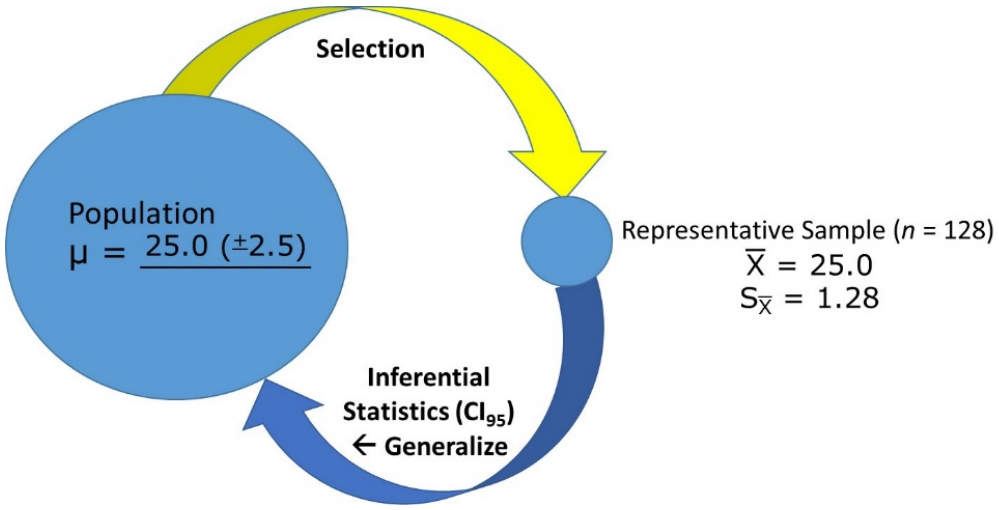

Figure 3. Correct use of inferential statistics requires both a population and a representative sample drawn from that population.

Inferential statistics are incorrectly used when either a census of the entire population is conducted or when a representative sample of the entire population is not studied. In the first instance, inference is not necessary because the researcher already has population parameters; there is no need (or benefit) to estimate a known population parameter. In the second instance, the researcher cannot generalize sample results back to a larger population using data obtained from a sample that does not represent that larger population.

In null hypothesis statistical testing (NHST), the p value is the conditional probability of obtaining the statistical test value if the null hypothesis is true (Spatz, 2018). The null hypothesis generally takes some form of the following:

> H_0: **In the population**, there is no difference in cognitive achievement between students taught by Method A, Method B, or Method C; or
> H_0: **In the population**, the correlation between Variable X and Variable Y is equal to zero (0); or
> H_0: **In the population**, the mean of Variable X is not equal to [hypothesized value].

Typically, in writing null hypotheses, researchers do not include the bolded text. However, it is included here to draw attention to the basic fact that all null hypotheses are hypotheses about the nature of reality in some population. Writing specifically about testing for group differences (such as represented by the first null hypothesis), Borg and Gall (1983) stated, "a test of statistical significance is made when we wish to determine how probable it is that the differences we found between our samples will also be found in the populations from which they were drawn" (p. 375). Note the specific linkage between samples and populations! If there is no population to which results can be generalized, there is no valid or logical reason to conduct or report the results of inferential statistical tests and their associated probabilities (Borg & Gall, 1983).

Making the correct decision on whether or not to use inferential statistics is simple (Figure 4). When you have data from a sample, representative of a population, and you wish to use the sample data to test hypotheses about the corresponding population parameter(s), select and use the correct inferential statistic. When you cannot meet both conditions, skip the inferential statistics and report population parameters or descriptive statistics, as appropriate (Miller, 1994).

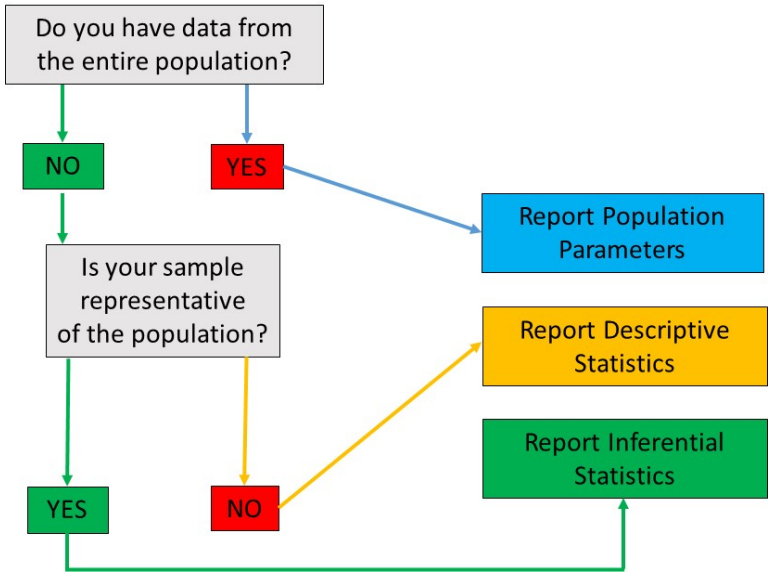

Figure 4. Decision tree for determining appropriate use of descriptive or inferential statistics.

Summary

Magic words and symbols are not adequate substitutes for rigorous thought, adherence to accepted standards, and careful and accurate writing when conducting and reporting quantitative agricultural education research. This manuscript has highlighted some areas for potential improvement in research reported by the profession. Following these suggestions will make it easier for our colleagues to 'peer over our shoulders' and make substantive judgements about the quality of our research and the reasonableness of the resulting conclusions and recommendations. Although none of us are perfect, continuous improvement toward this objective is a reasonable and worthwhile goal.

References

Ary, D., Jacobs, L., Sorenson, C. K., & Walker, D. (2014). *Introduction to research in education* (9th ed.). Ft. Worth, TX: Holt, Rinehart and Winston.

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

Borg, W. R., & Gall, M. D. (1983). *Educational research: An introduction* (4th ed.). New York City, NY: Longman.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York City, NY: John Wiley & Sons.

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

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*(2), 175-191.

Gates, H. R., Johnson, D. M., & Shoulders, C. W. (2018). Instrument validity in manuscripts published in the Journal of Agricultural Education between 2007 and 2016. *Journal of Agricultural Education, 59*(3), 185-197. doi: 10.5032/jae.2018.03185

Huck, S. W. (2008). *Reading statistics and research*. (5th ed.). Boston, MA: Pearson Education.

Israel, G. D. (2003). *Determining sample size*. Gainesville: Institute of Food and Agricultural Sciences, University of Florida, PEOD6.

Johnson, D. M., & Shoulders, C. W. (2018). Power of statistical tests used to address nonresponse error in the Journal of Agricultural Education. *Journal of Agricultural Education, 58*(1), 300-312. doi: 10.5032/jae.2017.01300

Kerlinger, F. N. (1986). *Foundations of behavioral research.* (3rd ed.). New York City, NY: Holt, Rinehart and Winston.

Linder, J. R., Murphy, T. H., & Briers, G. E. (2011). Handling nonresponse in social science research. *Journal of Agricultural Education, 42*(4), 43-51. doi: 10.5032/jae.2001.04043

McMillan, J. H., & Schumacher, S. (2010). *Research in educatio*n. (7th ed.). Upper Saddle River, NJ: Pearson Education.

Miller, L. E. (1994). Correlations: Description or inference? *Journal of Agricultural Education, 35*(1), 5-7. doi: 10.5032/jae.1994.01005

Miller, L. E. (2006). A philosophical framework for agricultural education research. *Journal of Agricultural Education, 47*(2), 106-117. doi: 10.5032/jae.2006.02106

Multon, K. D. (2010). Test-retest reliability. *Encyclopedia of Research Design*, Thousand Oaks, CA: SAGE Publications. doi: 10.4135/9781412961288

National Research Council. (1991). *In the mind's eye: Enhancing human performance*. Washington, DC: The National Academies Press. doi: 10.17226/1580.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrat4ed approach.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Qualtrics. (2018). *Sample size calculator*. Retrieved from https://www.qualtrics.com/blog/calculating-sample-size/

Reio, Jr. T. G. (2007). Survey nonresponse bias in social science research. *New Horizons in Adult Education and Human Resource Development, 21*(1/2), 58-51.

Rogelberg, S. G., & Stanton, J. M. (2007). Introduction: Understanding and dealing with organizational survey nonresponse. *Organizational Research Methods, 10*(2), 195-209. Retrieved from http://0-search.proquest.com.library.uark.edu/docview/195100352?accountid=8361Rogelberg

Spatz, C. (2019). *Exploring statistics: Tales of distributions.* (12th ed.). Conway, AR: Outcrop Publishers.

Stein, R., & Swan, A. B. (2019). Evaluating the validity of the Myers-Briggs type indicator theory: A teaching tool and window into intuitive psychology. *Social Personal Psychological Compass, 13*(3), 1-11. doi: 10.1111/spc3.12434

SurveyMonkey (n.d.). *Sample size calculator*. Retrieved from https://www.surveymonkey.com/mp/sample-size-calculator/

Warmbrod, R. J. (2014). Reporting and interpreting scores derived from Likert-type scales. *Journal of Agricultural Education, 55*(5), 30-47. doi:0.5032/jae.2014.05030