Journal of Agricultural Education,
56(3), 121-133. doi: 10.5032/jae.2015.03121

Peer Review in Agricultural Education: Interrater Reliability of Manuscript Reviews for the 2014 National Agricultural Education Research Conference

Catherine W. Shoulders[1], Donald M. Johnson[2], and Jim Flowers[3]

Abstract

This study analyzed 336 peer reviews of 112 manuscripts submitted for possible presentation at the 2014 National Agricultural Education Research Conference (NAERC). There were scoring errors on 6.8% of the reviews; the most frequent errors were failure to record a score or assigning a score above the range of points possible for one or more of the review criteria. The coefficient of variation (C_V) for same-paper ratings of manuscript quality ranged from 3.94% to 96.43% with a mean of 19.65% (SD = 15.22%). The interrater reliability for same-paper evaluations of manuscript quality was .15. The C_Vs for reviewer same-paper reject-accept recommendations ranged from 0.00% to 98.97% with a mean of 38.80% (SD = 22.72%). The interrater reliability for same-paper reject-accept recommendations was .09. On 82 (73.21%) manuscripts all three reviewers agreed on the manuscript's relevance to agricultural education. Mean manuscript quality scores explained 68.3% of the variance in mean reject-accept recommendations leaving 31.7% of the variance unexplained. There were no significant ($p > .05$) differences by academic rank in ratings of manuscript quality, reject-accept recommendations, or assessments of relevance to agricultural education. The authors offered recommendations for improvement of the NAERC manuscript review process.

Keywords: reliability, interrater, NAERC, manuscript review

In 1987, Camp, Hillison and Jeffries identified research productivity as the second most important factor (after faculty) in determining reputational quality rankings of university agricultural education programs. More than 20 years later, Birkenholz and Simonsen (2011) developed and tested a theoretical model of characteristics of distinguished agricultural education programs. As posited in their model, Birkenholz and Simonsen also found that research was one of the most often cited characteristics of distinguished programs.

Although agricultural educators publish in numerous journals and present at diverse conferences, the American Association for Agricultural Education (AAAE) sponsors three primary outlets for dissemination of scholarly work; the *Journal of Agricultural Education*, three regional research conferences, and the National Agricultural Education Research Conference (NAERC). Birkenholz and Simonsen (2011) identified professional meetings and presentations as one of the primary means through which faculty and departmental academic reputations are established and maintained. Agricultural education department heads rated presentation of papers at refereed

[1] Catherine W. Shoulders is an Assistant Professor of Agricultural Education in the Department of Agricultural Education, Communications and Technology at the University of Arkansas, 205 Agriculture Building, Fayetteville, AR 72701 (cshoulde@uark.edu).

[2] Donald M. Johnson is a Professor in the Department of Agricultural Education, Communications and Technology at the University of Arkansas, 205 Agriculture Building, Fayetteville, AR 72701 (dmjohnso@uark.edu).

[3] Jim Flowers is a Professor of Agricultural Education and Department Chair in the Department of Agricultural and Extension Education at North Carolina State University, Campus Box 7607, Raleigh, NC 27695 (jim_flowers@ncsu.edu).

research conferences as the second most important indicator of faculty research productivity; only publication of refereed journal articles was rated higher (Radakrishna & Jackson, 1993).

Papers to be considered for presentation at NAERC are limited to a maximum of 12 pages in length (single-spaced, 12-point Times New Roman font, with 1-inch margins) not counting the references (AAAE, 2014b). Each paper undergoes double-blind peer review by three reviewers using an evaluation rubric containing two scored components. The first component (manuscript quality) contains seven items for reviewers to evaluate overall writing quality and specific manuscript components (introduction, literature review and theoretical/conceptual framework, purpose and objectives, methods, results, and conclusions and recommendations). The manuscript quality evaluation component is internally weighted due to the different maximum point values assigned to different manuscript components (i.e. the introduction section has a maximum value of 5 points while the methods section has a maximum value of 15 points). The second component (reject-accept recommendation) contains a single item on a 1 to 6 Likert-type scale (1 = strongly disagree and 6 = strongly agree) asking reviewers if the paper should be presented at NAERC. A third un-scored component asks reviewers to assess (on a yes or no basis) if the manuscript is relevant to the agricultural education research agenda (Doerfert, 2011) or to agricultural education, broadly defined.

Raw scores for the two scored components (manuscript quality and reject-accept recommendations) are averaged across the three reviewers for each manuscript, converted to z-scores, weighted at .33 and .67, respectively, and added to determine each paper's final evaluation score. Based on AAAE's Protocol Guidelines for Conference Paper Selection, Presentations, and Awards (AAAE, 2011), the research conference chair may accept or reject a reviewed manuscript based on the three reject-accept recommendations alone or use the weighted score; there is no opportunity to revise and resubmit a rejected manuscript. The typical acceptance rate for NAERC is approximately 40% (Ricketts, 2013).

Given the importance of refereed conference presentations, especially NAERC presentations, in establishing faculty and departmental reputations (Birkenholz & Simonsen, 2011; Radakrishna and Jackson, 1993), a need existed to examine the peer review process by which papers are selected. Results from this study were expected to generate discussion and opportunities for improving the peer review process.

Conceptual Framework

Peer review is considered to be the cornerstone of scientific research (Baethge, Franklin, & Mertens, 2013). The British Academy (2007) stated, "Judgements about the worth or value of a piece of research should be made by those with demonstrated competence to make such a judgement" (p. 2). According to Daniel (1993), "Reviewers assume the role as 'Gatekeepers of Science'. . . recommending, in the ideal case, only those. . . manuscripts that meet the highest of scientific standards" (p. 1).

The journal *Philosophical Transactions* became the first peer-reviewed journal in 1752 (Spier, 2002) when the editor submitted manuscripts for "inspection by a select group of members who were knowledgeable in such matters, and whose recommendations were influential in the future progress of that manuscript" (p. 357). From this rather modest start, peer review has developed to the point where it is part of the fabric of scholarly inquiry (Hogden, 1997).

Yet, not all scholars have been entirely complimentary about the peer review process. In his forward to Daniel's (1993) classic, *Guardians of Science: Fairness and Reliability of Peer Review*, Nӗth wrote, "The Peer Review System. Some like it! Some dislike it! Some believe it is unfair! Some suspect it is ambiguous! Regardless of one's position, from the time of its inception in the 17th century it has remained controversial."

One of the oft-cited concerns about the peer review process is a perceived lack of agreement between reviewers as to manuscript quality (Bornmann, Weymuth, & Daniel, 2010; Daniel, 1993; Hodgson, 1997). According to Whitehurst (1984), "There are many reasons to be concerned with interrater agreement in peer review, not the least of which is a prevalent impression that the fate of a manuscript. . . is determined more by luck and editorial bias than to manuscript quality" (p. 26).

Pedhazur and Schmelkin (1991) defined reliability as "the ratio of true-score variance to observed-score variance" (p. 85). Under this definition, each manuscript review (an observed score) can be thought of as containing two components; a measure of actual manuscript quality (true score) and a combination of random and systematic errors (Shrout & Fliess, 1978). Thus, in the context of peer review of manuscripts, interrater reliability can be conceptualized as the extent to which two or more reviewers, using the same evaluation rubric, agree in their assessments of the quality and acceptability of the same manuscript (Huck, 2008; Kottner et al., 2011).

Several studies have been conducted evaluating the reliability of peer reviews. Marsh and Ball (1989) reported a mean interrater reliability of .27 between same-manuscript peer reviews for 10 social science journals. Kravitz et al. (2010) found an interrater reliability of .17 between same-manuscript reviews of 2264 manuscripts submitted to the *Journal of General Internal Medicine*. Hogden (1997) submitted the same grant proposal in the same funding year to two agencies that used the same proposal scoring system. The correlation between the two resulting peer review committee scores was .59, indicating that the scores from one panel explained only 35% of the variance in the scores from the other panel.

Studies have also focused on the qualifications of referees. Hamermesh (1994) concluded from a study which found that almost 12% of referees were from the same university department as the editors that referees may be appointed based on their relationship with the editor rather than on other qualifications. Stossel (1985) found that while the most highly-regarded individuals in a profession may be best suited to serve as referees, these same individuals have the greatest number of obligations, and therefore refuse to review papers more frequently than younger, more inexperienced colleagues. Other studies, however, have found the quality of reviews to be negatively correlated with referee seniority and status, causing the editors to actively seek out less experienced referees of lower academic status (Finke, 1990; Judson, 1994).

Purpose and Objectives

The purpose of this study was to evaluate the peer review of manuscripts submitted for presentation at the 2014 National Agricultural Education Research Conference. Specific objectives were to:

1. Determine the variability of same-manuscript ratings and the inter-rater reliability of peer reviews for manuscript quality scores;
2. Determine the variability of same-manuscript ratings and the inter-rater reliability of peer reviews for reject-accept recommendations;
3. Determine the level of same-manuscript agreement between peer reviewers on their assessment of each manuscript's relevance to agricultural education;
4. Determine the relationships between manuscript quality scores, reject-accept recommendations, and assessment of relevance to agricultural education;
5. Determine whether manuscript quality scores, reject-accept recommendations, and assessments of relevance to agricultural education differed by the academic rank (instructor, assistant professor, associate professor or professor) of the reviewer.

Methods

After institutional IRB approval and approval by the AAAE Research Committee, the lead researcher was provided with access to the Fast Track® manuscript review files for each of the 112 manuscripts submitted for the 2014 NAERC. Each manuscript had been reviewed by three peer reviewers for a total of 336 reviews. Review data were manually entered into an Excel spreadsheet and each reviewer's academic rank was determined using either the AAAE Directory (AAAE, 2014a) or the reviewer's institutional web site. To ensure reviewer anonymity, no reviewer identification information other than academic rank was included in the data set. All data were verified and imported into SAS (Version 9.3) for data analysis.

All reviewers completed their NAERC manuscript evaluations using an online rubric. The rubric contained seven quantitative items for evaluating manuscript quality; introduction (1 - 5 points), literature review and conceptual/theoretical framework (1 - 11 points), purpose and objectives (1 -7 points), methods/procedures (1 - 15 points), results (1 - 11 points), conclusions and recommendations (1 - 15 points), and writing quality (1 - 5 points); total manuscript quality scores could range from 7 to 69 points. The reject-accept recommendation item asked reviewers to rate their level of agreement, on a 1 to 6 scale (1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = slightly agree, 5 = agree, and 6 = strongly agree), that the manuscript should be accepted for presentation at the conference. A single yes or no item asked whether the study was clearly linked to the agricultural education research agenda or to agricultural education broadly defined.

Objectives one and two sought to determine the variability and interrater reliability of manuscript quality scores and reject-accept recommendations. For both objectives, the mean total score, standard deviation and coefficient of variation (C_V) were calculated on the reviews for each manuscript and their distributions were examined. The C_V, which expresses the standard deviation as a percentage of the mean ($C_V = M / SD$ x 100), was used as a descriptive measure of the relative variation of reviewer scores for the same manuscript (Freund & Wilson, 1993). According to Abdi (2010), the C_V allows direct comparison of the variation in two or more variables (manuscript reviews) with different means or measured on different scales (such as manuscript quality scores and reject-accept recommendations). Higher C_Vs indicate greater variation among reviewers' ratings.

Intraclass correlation coefficients (ICC) were calculated as the measure of interrater reliability (Shrout and Fliess, 1979). ICC coefficients range from -1/*c* (where *c* is the number of reviewers per manuscript) to 1.00 (Whitehurst, 1984); since the 2014 NAERC used three reviewers per manuscript, the lower limit on negative scores for this study was -.33. As a practical matter, negative coefficients are interpreted as zeros while values of 1.0 are interpreted as perfect agreement among reviewers (Whitehurst, 1984). The ICC "provides an index of the reliability of the ratings for a *single, typical* judge" (MacLennan, 1993, p. 294) and is the most appropriate measure of interrater reliability (Cicchetti, 1980).

According to Whitehurst (1984), the use of ICC for manuscript reviews "corresponds to a one-way random effects analysis of variance (ANOVA) in which separate and independent ratings of each manuscript are treated as subjects" (p. 81). In this analysis, "the intraclass correlation takes the form of the ratio of the variance attributable to manuscripts to the variance attributable to manuscripts plus the error components" (p. 23). This is consistent with the definition of reliability as the "the ratio of true-score variance to observed-score variance" (Pedhazur & Schmelkin, 1991, p. 85).

Shrout and Fliess (1979) identified six forms of ICC depending on the statistical model of the reliability study. In this study, each manuscript was evaluated by a different set of peer reviewers and each scored component (quality score and reject-accept recommendation) was considered a separate item. Thus, the researchers identified the ICCs for objectives one and two as meeting the requirements of ICC(1,1) as described by Shrout and Fliess (1979).

Objective three sought to determine the level of agreement between reviewers on the relevance of the manuscript to agricultural education. Since each reviewer rated this on a yes or no basis, agreement percentages were calculated (Kottner et al., 2011). Although agreement percentages have been criticized for failing to correct for agreement due to chance (Lombard, Snyder-Duch, & Bracken, 2000), the researchers selected this method because, on a practical level, the actual level of agreement, regardless of its source, was the variable of interest. From a research perspective, correcting for chance agreement results in a statistical model of agreement instead of a description of agreement (Uebersax, 1992).

Objective four sought to determine the relationships between manuscript quality scores, reject-accept recommendations, and assessment of the manuscripts' relevance to agricultural education. The magnitude of each correlation was described using the descriptors proposed by Davis (1971).

The final objective sought to determine if differences existed in manuscript quality scores, reject-accept recommendations, or assessments of relevance to agricultural education by reviewer academic rank (instructor or similar, assistant professor, associate professor or professor). One-way analysis of variance (ANOVA) and the chi square test of association were used to analyze data for this objective.

For objectives one through four, manuscripts (N = 112) were the unit of analysis and all statistical results are based on the three reviews for each manuscript. For objective five, individual manuscript reviews (N = 336) were the unit of analysis; no attempt was made to correct for non-independence between observations caused by reviewers evaluating multiple manuscripts.

Reviews from the 2014 NAERC were considered to be a time-place sample representative of past and future NAERC manuscript reviews. As such, the use of inferential statistics was warranted (Oliver & Hinkle, 1982). The .05 alpha level was selected *a priori* for all tests of statistical significance.

Results

One hundred and twelve manuscript submissions were each reviewed by three reviewers for the 2014 NAERC, for a total of 336 manuscript reviews conducted by 151 individual peer reviewers. Of the 336 manuscript reviews conducted, 23 (6.8%) contained one or more scoring errors. The most frequent errors were failing to record a score for an evaluation criterion (18 occurrences) and assigning a score above the rubric's maximum value for an evaluation criterion (16 occurrences). Conference managers typically contact reviewers to resolve errors in scoring, but these corrections cannot be edited in the reviewer's submitted manuscript evaluation rubric (D. Doerfert, personal communication, August 15, 2014). Because the purpose of this study was to determine the reliability of the peer review process as conducted by the peer reviewers, all 336 reviews were included in the analysis.

Objective 1

Objective 1 sought to determine the variability of same-manuscript ratings and the inter-rater reliability of peer reviews for manuscript quality scores. The distribution of the mean quality scores for each manuscript (based on three reviews per manuscript) was negatively skewed (skewness = -0.66) and ranged from 28.00 to 64.33 with a grand mean of 50.24 (SD = 7.47). The distribution of within-manuscript standard deviations was positively skewed (skewness = 1.22) and ranged from 1.73 to 30.99 with an overall mean of 9.18 (SD = 5.48).
The distribution of C_Vs for manuscript quality scores (Figure 1) was positively skewed (skewness = 2.34) and ranged from 3.94% to 96.43% with a mean of 19.65% (SD = 15.22%). None of the 112 manuscripts had perfect agreement between the three reviewers as to manuscript quality score. The interrater reliability of manuscript quality scores for the 2014 NAERC was .15.

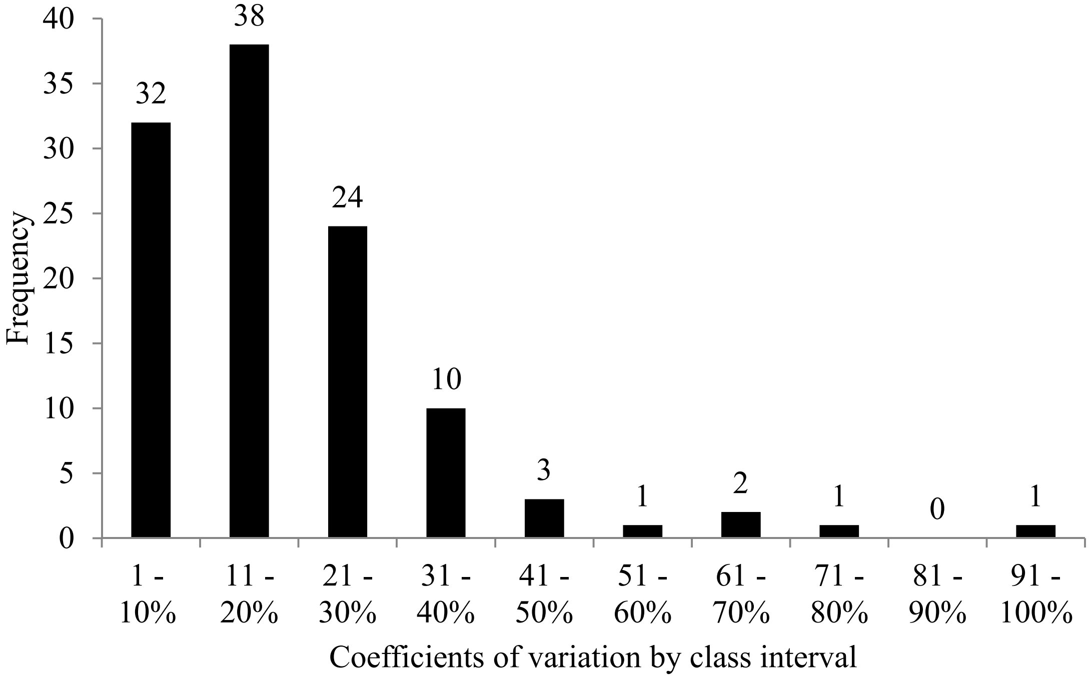

Figure 1. Distribution of coefficients of variation for within-manuscript quality scores.

Objective 2

Objective 2 sought to determine the variability of same-manuscript ratings and the inter-rater reliability of peer reviews on reject-accept recommendations. The distribution of mean reject-accept recommendations for each manuscript (based on three reviews per manuscript) had a slight negative skew (skewness = -0.14) and ranged from 1.33 to 5.67 with a grand mean of 3.98 (*SD* = 1.01). The distribution of within-manuscript standard deviations exhibited negligible skewness (0.02) and ranged from 0.00 to 2.89 with a mean of 1.39 (*SD* = 0.71). The distribution of C_Vs for manuscript reject-accept recommendations had a slight positive skew (skewness = 0.21) and ranged from 0.00% to 98.97% with a mean of 38.80% (*SD* = 22.72%). As shown in Figure 2, raters were in perfect agreement in their reject-accept recommendations for 5 of the 112 (4.46%) manuscripts; for 35 (31.25%) manuscripts the C_V was more than 50%. The interrater reliability for accept-reject recommendations for 2014 NAERC manuscripts was .09.

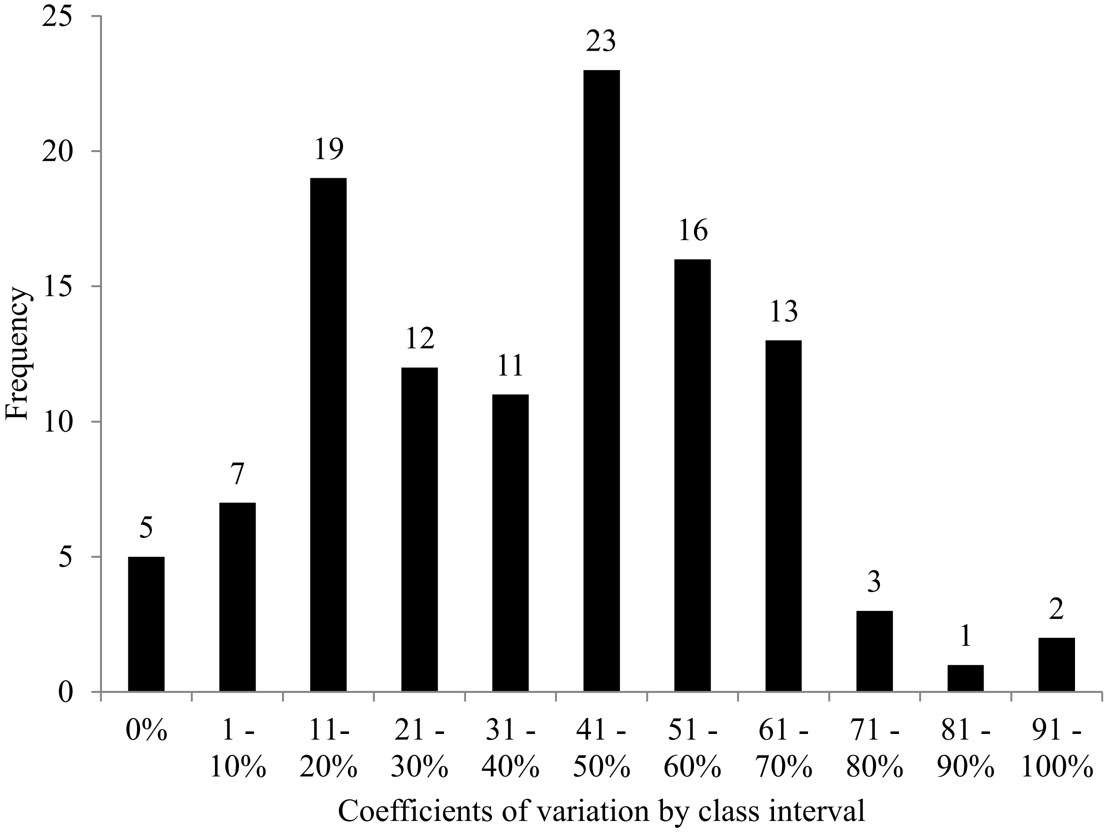

Figure 2. Distribution of coefficients of variation for within-manuscript reject-accept recommendations.

Objective 3

Objective 3 sought to determine the level of same-manuscript agreement between peer reviewers on their assessment of each manuscript's relevance to agricultural education. A single dichotomous item on the evaluation rubric asked reviewers to indicate whether or not the author(s) had clearly linked the study to an agricultural education priority research area (Doerfert, 2011) or to an area of agricultural education not specifically identified in the research agenda. On 82 (73.21%) manuscripts all three reviewers agreed this linkage had been made; on 20 (17.86%) manuscripts two reviewers agreed and one reviewer disagreed; on seven (6.25%) manuscripts two reviewers disagreed and one reviewer agreed. The remaining three (2.68%) manuscripts each had one missing reviewer response to this item.

Objective 4

Objective 4 sought to determine the relationships between manuscript quality scores, reject-accept recommendations, and assessment of relevance to agricultural education. There was a significant ($p < .0001$) and very strong (Davis, 1971) positive correlation ($r = .83$) between the mean quality scores and the mean reject-accept recommendations for the 112 manuscripts submitted for the 2014 NAERC. Mean quality scores explained 68.3% of the variance in mean reject-accept recommendations; however, 31.7% of this variance was not explained by mean quality scores.

There was a significant ($p = .02$) and low (Davis, 1971) positive correlation ($r_s = .24$) between mean relevance ratings and mean reject-accept decisions for the 112 manuscripts. The correlation ($r_s = .16$) between mean relevance ratings and mean manuscript quality scores was not statistically significant ($p = .08$).

Objective 5

Objective 5 sought to determine whether differences existed in manuscript quality scores, reject-accept recommendations, and assessment of relevance to agricultural education by reviewer academic rank. One-hundred-forty-seven unique reviewers performed 336 reviews. Thirteen reviewers (8.84%) were instructors or similar, 76 (51.70%) were assistant professors, 24 (16.33%) were associate professors, and 34 (23.13%) were professors. Of the 336 individual reviews conducted on the 112 manuscripts submitted for the 2104 *NAERC*, 47.6% were completed by assistant professors, 22.2% by professors, 21.6% by associate professors and 8.7% by instructors or similar reviewers. One-way ANOVAs revealed no significant differences ($p > .05$) by reviewer rank in manuscript quality scores, $F(3, 329) = 0.18$, $p = .91$, or in reject-accept recommendations, $F(3, 329) = 0.68$, $p = .57$. Finally, chi square analysis found no significant ($p > .05$) difference by rank in the percentage of reviewers agreeing or disagreeing about the manuscripts' relevance to agricultural education, $\chi^2(3) = 5.51$, $p = .14$. Table 1 provides descriptive statistics for these variables by reviewer rank.

Table 1

Descriptive Statistics for Manuscript Quality Scores, Presentation Recommendation Scores, and Relevance to Agricultural Education, by Reviewer Academic Rank

		Quality Score		Reject-Accept Recommendation		Relevant to Agricultural Education?	
Reviewer Rank	*N*	*M*[a]	*SD*	*M*[b]	*SD*	Yes (%)	No (%)
Instructor or Similar	27	51.46	10.55	4.21	1.40	89.29	10.71
Assistant Professor	159	50.33	11.01	3.96	1.58	85.53	14.47
Associate Professor	72	49.64	11.54	3.79	1.71	92.86	7.14
Professor	75	50.03	12.94	4.11	1.71	94.52	5.48

[a]Valid range of possible quality scores was 7 to 69; due to scoring errors, actual scores ranged from 1 to 71. [b]Based on a 6-point scale (1 = "strongly disagree" and 6 = "strongly agree.")

Conclusions, Discussion and Recommendations

A small but critical percentage (6.8%) of all manuscript reviews contained one or more scoring errors. The most common errors were failure to record a score or recording a score above the maximum value for one or more manuscript quality criteria. These errors may be the result of reviewers' attempts to reduce the time required to conduct reviews, as the responsibilities of faculty members are numerous. Conference managers have the responsibility of identifying these errors and contacting reviewers to make corrections, adding to their workload.. Because the currently employed Fast Track® manuscript submission and review system does not allow automatic data validation or required-response features (D. Doerfert, personal communication, August 15, 2014), the researchers recommend that the AAAE Research Committee explore alternative manuscript submission and review systems that use technology to reduce potentially unnecessary editorial responsibilities in order to reduce the burden of manuscript scoring on both reviewers and the conference manager.

The distribution of C_Vs for same-manuscript quality scores ranged from 3.94% to 96.43% with a mean of 19.65%. Thus, for the average manuscript, the standard deviation between the three reviewers' quality ratings was nearly 20% as large as the mean manuscript quality score. A majority of C_Vs were clustered fairly closely around the mean. However, a significant subset of higher C_Vs was observed, resulting in a positively skewed distribution. This subset indicates that while the majority of the reviewers produced similar scores for manuscripts, a small number of reviewers differed greatly in the scores they assigned to a manuscript. This subset of manuscripts was also impactful on the interrater reliability of manuscript quality scores, which was .15. This is similar to the .17 value reported for the *Journal of General Internal Medicine* (Kravitz et al., 2010) but lower than the .27 reported as the mean of 10 social science journals (Marsh & Ball, 1989). To fully display the impact a small number of manuscript scores had on interrater reliability, 2014 NAERC manuscript quality scores were reanalyzed after removing the 28 manuscripts with C_Vs above the 75th percentile (C_V > 24.53%). The interrater reliability increased (from .15) to .54 for reviews of the 84 remaining manuscripts.

The profession should engage in thought and dialogue about what constitutes quality in each of the seven components evaluated, as well as how to educate reviewers in determining that quality. Reviewer workshops, including sample reviews with subsequent discussion, should be conducted at the regional and national AAAE conferences. While total agreement on quality scores is neither likely nor desirable, questions reviewers may have regarding the intentions and impact of scoring components may be addressed during these workshops, thereby improving reviewers' ownership and confidence in the review process.

The distribution of C_Vs for reject-accept recommendations ranged from 0.00% to 98.97% with a mean of 38.80%. Thus, for the average manuscript, the standard deviation between the three reviewers' reject-accept recommendation was nearly 40% as large as the mean manuscript reject-accept score. The mean C_V for reject-accept recommendations was 1.97 times as large as the mean C_V for quality scores. The interrater reliability of reject-accept recommendations was .09, suggesting that reviewers are less consistent in their reject-accept recommendations than they are with manuscript quality scores. Mean manuscript quality scores explained 68.3% of the variance in mean reject-accept recommendations. Further research is warranted to determine what subjective factors account for the 31.7% of variance not explained by manuscript quality.
As with manuscript quality scores, the interrater reliability of reject-accept recommendations was recalculated after removing the 28 manuscripts with C_Vs above the 75th percentile (C_V > 57.51%). The interrater reliability increased (from .09) to .23 for reviews of the remaining 84 manuscripts. Thus, reviewer training may not impact the interrater reliability of reject-accept recommendations given that reviewers may have their own idiosyncratic notions, independent of quality, concerning what constitutes an "acceptable" NAERC manuscript.

Theory holds that the reliability of a measurement scale increases as the number of items increases (Pedhazur & Schmelkin, 1991). Thus, weighting of the single-item reject-accept recommendation as 67% of the manuscript decision process should be reevaluated. Opportunities for adjustment may include discontinuing use of this item, reducing its weighting, or redesigning it as a multi-item scale. Additionally, discussion within the profession regarding the purpose of these two related but separately-weighted review criteria may assist in the reduction of the time and effort required of manuscript reviews. If the profession agrees that all variation in reject-accept decisions shouldbe attributed to manuscript quality score, the need for the additional reject-accept criterion may be eliminated, thus reducing the length of the manuscript review. However, if the profession determines that variation within reject-accept decisions is associated with evaluation components other than quality score, consensus must be reached as to what factors referees should be using to assign reliable reject-accept scores, thereby reducing the burden of subjectivity on the reviewer.

Reviewers disagreed on whether or not the author(s) had linked their manuscripts to the agricultural education research agenda on 26.79% of same-manuscript reviews. Since this item is not scored for manuscript selection and had a no significant relationship to manuscript quality scores, the researchers encourage the AAAE Research Committee to consider either deleting this item from the evaluation rubric or incorporating it into one of the evaluation sections in the manuscript score. Again, removal of this item can reduce the length of the manuscript review process, reducing the burden on reviewers. Although the researchers are supportive of the intent of the item, its inclusion as a stand-alone item does not appear to serve a valid evaluative purpose.

Finally, reviewer academic rank had no significant relationship to manuscript quality scores, reject-accept recommendations or assessments of a manuscript's relevance to agricultural education. Thus, the probability a manuscript will be accepted or rejected is not affected by the academic rank of the reviewers, suggesting that the quality of the profession's referee process is neither hindered nor strengthened by the use of higher- or lower-ranked referees. Conference planners should continue seeking reviews from all members of the profession, regardless of academic status.

For the 2014 NAERC, manuscripts were reviewed by instructors or similar, 8.0%; assistant professors, 47.3%; associate professors, 21.4%; and professors, 22.3%. When compared to the number of unique reviewers within each academic rank, the percentage of unique reviewers at the associate professor level was lower (16.33%) than the percentage of reviews conducted overall by associate professors (21.60%). The manager of the conference takes on the responsibility of performing reviews either declined or not completed by other reviewers; in 2014, the conference manager was an associate professor. His unprecedented high number of reviews is likely to have inflated the number of reviews taken on by the associate professor ranks within the profession, suggesting that the overall group has performed a higher number of reviews, when in fact, one associate professor has performed a higher number of reviews. There were the greatest number of unique assistant professors serving as reviewers; the percentage of unique reviewers at the assistant professor level (51.70%) greatly exceeds their percentage within the professional membership (29.9%). According to the most recent data available (Swortzel, 2011), the AAAE faculty membership consists of 10.8% instructors, 29.9% assistant professors, 24.6% associate professors, and 34.7% professors. Clearly assistant professors, who made approximately half of the unique reviewers and conducted almost half of the reviews, but only make up 29.9% of the membership population, are carrying a disproportionate share of the professional load in reviewing NAERC manuscripts, as has been previously posited by Stossel (1988). Associate professors and, professors should increase their level of involvement in reviewing NAERC manuscripts when time and availability allow in order to enable assistant professors to focus more fully on developing all aspects of scholarship before being considered for tenure and promotion. Further, the unwritten research expectations and standards within a profession are unfamiliar to newer faculty members; distinguished faculty members have an opportunity to pass down a profession's research traditions through feedback during the review process, thereby maintaining the integrity of the profession's research for future generations.

As demonstrated by the literature, an ideal peer-review process has yet to be found in any research profession. This study sought to describe the current status of the NAERC peer review process and bring to light both strengths and weaknesses of the process with the intention of generating discussion among members of the profession. The results indicated the reliability of 2014 NAERC manuscript reviews were low but similar to those of some refereed journals. Periodic evaluation of a discipline's peer-review process enables the discipline to identify and address components of the process needing improvement. As the agricultural education profession continues to improve its peer-review process, this study can serve as a benchmark on which to evaluate the results of these efforts.

References

AAAE. (2009). AAAE protocol guidelines for conference paper selection, presentations and awards. Columbus, OH: American Association for Agricultural Education.

AAAE. (2014a.) Ag education directory. Available at http://aaaeonline.org/directory_query _select.php

AAAE. (2014b). Call for research papers, American Association for Agricultural Education 2014 annual conference. Retrieved from American Association for Agricultural Education website: http://aaaeonline.org/uploads/allconferences/8-27-2013_146_NAERC_2014_Call_for_Papers_copy.pdf

Abdi, H. (2010). Coefficient of variation. In Neil J. Salkind (Ed.), *Encyclopedia of Research Design.* (pp. 170-172). Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781412961288.n56

Baethge, C., Franklin, J. & Mertens, S. (2013). Substantial agreement of referee recommendations at a general medical journal - a peer review evaluation at *Deutsches Ärzteblatt International*. *PLOS One, 8*(5), 1-7. Retrieved from http://europepmc.org/articles/PMC3642182

Birkenholz, R. J., & Simonsen, J. C. (2011). Characteristics of distinguished programs of agricultural education. *Journal of Agricultural Education,52*(3), 16-26. doi: 10.5032/jae.2011.03016

Bornmann, L., Mutz, R., Daniel, H. D. (2010). A review-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLOS One, 5*(12), 1-10. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0014331

British Academy (2007). *Peer review: The challenges for the humanities and social sciences.* London: The Academy.

Camp, W. G., Hillison, J., & Jeffries, B. J. (1987). Peer rankings of the leading agricultural teacher education programs. *Journal of the American Association of Teacher Educators in Agriculture, 28*(4), 2-8. doi:10.5032/jaatea.1987.04002

Cicchetti, D. V. (1980). Reliability of reviews for the *American Psychologist*: A biostatistical assessment of the data. *American Psychologist, 35*(3), 300-303.

Daniel, H. D. (1993). *Guardians of science: Fairness and reliability in peer review.* New York: VCH Publishers.

Davis, J. A. (1971). *Elementary survey analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Doerfert, D. L. (Ed.) (2011). *National research agenda: American Association for Agricultural Education's research priority areas for 2011-2015*. Lubbock: Texas Tech University, Department of Agricultural Education and Communications.

Finke, R. A. (1990). Recommendations for contemporary editorial practices. *American Psychologist, 45*, 669-670.

Freund, R. J., & Wilson, W. J. (1993). *Statistical methods.* San Diego, CA: Academic Press.

Hamermesh, D. S. (1994). Facts and myths about refereeing. *Journal of Economic Percpectives, 8*, 153-163.

Hogdon, C. (1997). How reliable is the peer review? An examination of operating grant proposals simultaneously submitted to two similar peer review systems. *Journal of Clinical Epidemiology, 50*(11), 1189-1195.

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

Judson, H. F. (1994). Structural transformations of the sciences and the end of peer review. *Journal of the American Medical Association, 272*, 92-94.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A.,... Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International Journal of Nursing Studies, 48*(2011), 661-671. dio: 10.1016/j.inurstu.2011.01.016.

Kravitz, R. L., Franks, P., Feldman, M. D., Gerrity, M., Byrne, C., & Tierney W. M. (2010). Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care? *PLOS One, 5*(4), 1 -5. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0010072

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication. *Human Communication Research, 28*(4), 587-604.

MacLennan, R. N. (1993). Interrater reliability with SPSS for Windows 5.0. *The American Statistician, 47*(4), 292-296.

Nöth, H. (1993). Foreward. In H. D. Daniel, *Guardians of science: Fairness and reliability of peer review*. New York: VCH Publishers.

Oliver, J. D., & Hinkle, D. E. (1982). Occupational education research: Selecting statistical procedures. *Journal of Studies in Technical Careers, 9*(1982), 199-207.

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

Radhakrishna, R. B., & Jackson, G. (1993). Agricultural and extension education department heads' perceptions of journals and the importance of publishing. *Journal of Agricultural Education, 34*(4), 8-16. doi: 10.5032/jae.1993.04008.

Rickets, C. (2013). AAAE submission & review manager's report. Retrieved from American Association for Agricultural Education website: http://aaaeonline.org/allconferences1.php?show_what=National&sorter_conf=National&sorter_year=2013

Shrout, P. E., & Fliess, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Spier, R. (2002). The history of the peer review process. *TRENDS in Biotechnology, 20*(8), 357-358.

Stossel, T. P. (1985). Reviewer status and review quality: Experience of the *Journal of Clinical Investigation. New England Journal of Medicine, 312*, 658-659.

Swortzel, K. A. (2011). *American Association for Agricultural Education 2010 - 2011 faculty salary report.* Retrieved from American Association for Agricultural Education website: http://aaaeonline.org/files/salary/salary10-11.pdf

Uebersax, J. S. (1992). Modeling approaches for the analysis of observer agreement. *Investigative Radiology, 27*(9), 738-743.

Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist, 39*(1), 22-28.