jmde

http://evaluation.wmich.edu/jmde/ *Articles*

Peer Review of Submissions to the Annual American Evaluation Association Conference by the Graduate Student & New Evaluators Topical Interest Group

Daniela C. Schröter
The Evaluation Center, Western Michigan University

Chris L. S. Coryn
The Evaluation Center, Western Michigan University

Bianca E. Montrosse
The SERVE Center, University of North Carolina at Greensboro

Peer review is an umbrella term that refers to a class of selection and oversight practices, including the familiar mechanisms of the review of proposals submitted for funding, of manuscripts for scholarly publications, and of personnel qualifications and portfolios for selection and promotion. Peer review has long been a cornerstone of modern scientific method premised on the assumption that those within a discipline are best suited to assess the work of others within that field. As such, it is also frequently employed to evaluate proposals submitted for professional meetings such as the annual conference of the American Evaluation Association (AEA). This paper presents a blind peer review method developed by AEA's Graduate Student & New Evaluators (GS&NE) Topical Interest Group (TIG) in an effort to construct an impartial and reliable process in proposal selection. Implications for conference review processes, AEA, and the field of evaluation in general are discussed.

Peer review is the name given to judgments of scientific merit by other scientists working in, or close to the field in question. Peer review is premised upon the assumption that a judgment about certain aspects of science, for example, its quality, is an expert decision capable of being made only by those who are sufficiently knowledgeable about the cognitive development of the field, its research agendas, and the practitioners within it (OECD 1987).

Peer review is a blanket term for a family of selection and oversight practices, including the well-known systems employed for the evaluation of grant proposals, manuscript review for scholarly journals, and research and personnel evaluation, among others (Coryn, 2007). In essence, peer review as discussed in this paper is one form of proposal evaluation that has been defined as one branch of evaluation (Coryn & Hattie, 2006; Scriven, 2003). Although the mechanics of peer review are generally familiar, they are also complex and idiosyncratic. Too often, discussions of peer review narrowly focus on technical matters such as interrater agreement, conflicts of interest, normalization of raters' scores to achieve comparability across reviewers, and other problems associated with process (Hackett, 1997; Foltz, 2000).

A Brief History of Peer Review

The origins of contemporary peer review can be traced back to the emergence of scientific journals in the early 17th century (Langfeldt, 2002; Pyenson & Sheets-Pyenson, 1999). However, as a cornerstone of modern scientific method, peer review has only been consistently applied since the middle of the 20th century (Aksnes, 2005). Prior to World War II, there were no universally adopted standards or norms for evaluating scientific research. Practices were conducted independently by each journal in response to idiosyncratic conditions (Burnham, 1992). For instance, Albert Einstein's influential *Annus Mirabilis* papers (1905a, 1905b, 1905c, 1905d, 1905e), which appeared in *Annalen der Physik* were not peer-reviewed. In fact, the journal's editors Max Planck and Wilhelm Wien merely recognized the virtue of such innovative ideas and simply published the papers.

In the post-World War II era, peer review practices have become increasingly more sophisticated and systematic with the introduction of double-blind and single-blind—as opposed to open review—procedures as quality controls for disseminating research (Campanario, 1998). The double-blind process is one where not only the referees remain anonymous to the authors, but where the authors also remain anonymous to the referees, whereas single-blind procedures are where the reviewer knows the identity of the author but not vice versa (Justice, Cho, Winker, Berlin, & Rennie, 1998; Mainguy, Motamedi, & Mietchen, 2005; McNutt, Evans, Fletcher, & Fletcher, 1990). In open peer review, the identities of both authors and reviewers are revealed, affording the authors the ability to identify the reviewers (van Rooyen, Godlee, Evans, Black, & Smith, 1999; Walsh, Rooney, Appleby, & Wilkinson, 2000). The main argument against open peer review is that junior reviewers will be reluctant to criticize the work of senior researchers for fear of reprisals. This fear is particularly acute for researchers whose livelihoods depend on winning grants (Smith, 1999). The principal argument in favor of blinding is that:

> …the signing of reviews would inhibit reviewers from being open and probing in their critiques (as has increasingly happened with letters of personal recommendation); this would clearly not be in the best interests of good science. The principal argument against blinding is that it might foster irresponsibility, particularly slanted and destructive criticism, because reviewers know that authors cannot hold them personally accountable for their opinions. The case for "opening up" peer review by identifying reviewers to authors is therefore being vigorously put forward (Davidoff, 1998, p. 66).

By and large, peer review is almost universally the predominant method used for evaluating research; it is seen as an obligatory system within the scientific community and widely perceived as the only legitimate method for valuing scientific merit (Coryn, Hattie, Scriven, & Hartmann, 2007; Coryn & Scriven, 2008). Nonetheless, it is claimed that peer review is "partial, biased and unreliable, and it takes time away from research activities" (Langfeldt 2002, p. 16). For instance, it is more or less directly claimed that the peer review system is essentially an "old boys' club," which is full of scientists feathering their own nests, favoring eminent scientists (i.e., the *halo effect*), and stifling innovative research because assessments are done by well-established researchers who reject ideas that differ from their own. That is, the system often discriminates against scientists who work in "low-prestige" institutions and is sometimes punitive against innovation (Cole, Cole, & Simon, 1980; Coryn, 2006; Turney 1990; McCook, 2006). Additionally, peer review has been characterized as

> …unreliable, fashion-based, and manipulable … [but] … like democracy, peer review may be a flawed system, but if given its best possible implementation, it is the best in sight and something like it will always be a key element in

> proposal and program evaluation … [and] … we need to correct its imperfections, and knowledge about its weaknesses is the best place to start planning improvements (Scriven, 1993, p. 86).

Typical Features of Peer Review

Because peer review has been used to evaluate individual researchers or research products for decisions about employment, promotion, publication, and funding traditionally, empirical studies of peer review processes as applied to the evaluation of abstracts submitted for presentation at professional meetings and conferences are virtually nonexistent. Nevertheless, the underlying ideal is comparable to that of peer reviews conducted for journal publication. Although the stakes are considerably lower in evaluating conference proposals than in decisions regarding publication, the rationale remains the same—promoting high quality work that represents the full range of good and original ideas among the members of an academic or professional community.

Typically, peer review for journal publication can be characterized as a process where editors use systematic procedures to distribute the work to expert reviewers for evaluation and seek consensus about the quality of the work to make a decision as to whether the submission should be accepted, rejected, or revised (Benos et al., 2007). Alternatively, the review process for professional conferences, such as AEA, can be characterized as a process where the conference management systematically distribute submissions to respective Topical Interest Group (TIG) program chairs (who function as the editor for their TIG) who involve expert and non-expert reviewers, in most cases, to systematically and objectively evaluate conference submissions based on predetermined criteria (set by AEA or developed by the TIG), and who assess the merit of the submission for presentation at the annual conference (Schröter, 2007). Decisions are then made about acceptance or rejection.

Although the peer review process for publication has similarities to those used for conference reviews, there are also fundamental differences. For example, it is uncommon that a submission for conference presentation can be improved, revised, and resubmitted (McEneaney, 2001). Although open review systems would allow for such improvements, time constraints, variation in submission formats, and other issues associated with conference planning do not necessarily support such practices. In contrast to submissions to journals, a decision about a submission for conference presentation is usually final, at least for the given year. Either a proposed presentation is accepted, or it is not.

Therefore, peer review in this context is normally summative in nature. However, it can also be formative or, less frequently, ascriptive. Unlike formative and summative peer review evaluations, ascriptive evaluation is neither aimed at improvement nor at decision making, specifically, and is normally done merely for the sake of knowing. That is, peer review done for ascriptive purposes is roughly equivalent to Patton's (1997) and Chelimsky's (1997) notion of evaluation's function to generate knowledge. Ascriptive types of peer review evaluation are those often conducted as part of the day-to-day process of the scientific endeavor, for example, assessing the quality of the previous literature or the explanatory power of a theory. Furthermore, the formative and summative roles of peer review are not always mutually exclusive, and are occasionally orthogonal. Nevertheless, one thing is clear, the logic and lexicography of evaluation "does require that both formative and summative evaluation involve efforts to determine merit" (Scriven, 1996, p. 157) and that this distinction is ultimately context dependent. For instance,

> …an editorial decision to "accept" or "reject" a research manuscript submitted for publication is summative, while a decision of "revise and resubmit" is formative. For the author, however, both the reject and revise and resubmit

> decisions can be formative in that the author can opt to improve the manuscript, especially if feedback was given by reviewers or the editors. In any case, the decisions are *de facto summative* [italics in original] in the editorial context, but nearly always formative in the context of the author. Of course, the author could simply make a decision to submit the manuscript to another journal, in which case the author has undertaken a summative evaluation of another kind—a decision not to make use of the editors' recommendations (i.e., revise and resubmit) as a basis for improvement (Coryn, 2007, p. 36).

In any case, the problems of peer review for both publication in journals or presentation at conferences are largely the same, and include "concerns … about bias, fairness, unnecessary delay, and general ineffectiveness of the process" (Benos et al., 2007, p.145), among others.

Bias and fairness in peer review are related to favoritism of reputable personae or institutions, gender bias, differences in ideologies, and general conflicts of interest (Ceci and Peter, 1982; Ross et al., 2006). Ross et al. (2006) found that reviewer bias (as presented in reviews conducted for the American Heart Association's annual Scientific Sessions) were partially reduced by blinding abstracts from authors' names and institutional affiliations.

Delay is common in peer review for journal publication. Editors wait for reviewer comments, authors revise and resubmit their articles, articles are re-reviewed, and by the time the article appears, it is often outdated (i.e., publication lag). A paper in educational technology suggests that reviews in this field last between two to 24 weeks, while reviewers typically receive four to six weeks for reviewing manuscripts (Niederhauser, Wetzel, & Lindstrom, 2004). In reviews for conference proposals, the peers do not have that much time; revisions are not an option and those involved in the review process are tied to the conference schedule. In the end, those who coordinate the review process must find strategies to yield decisions in a timely manner, so that conference planning can proceed. For example, in 2007 AEA TIGs had one month to complete the review of all submissions to their TIGs. Within this timeframe, TIG program chairs spent an average of 17 hours (SD = 22.56; one TIG reported 130 hours) to implement and complete the review process (Schröter, 2007).

Focus of This Article

This article discusses a peer review process developed by AEA's Graduate Student and New Evaluator (GS&NE) TIG in 2006. To date, there has been no contribution about review processes used by the evaluation community to assess submissions for presentation at annual conferences, such as AEA, and other contributions about peer review for conferences or professional meetings are rare. We believe this article to be of importance to all those readers who regularly submit proposals or who are involved in review processes for conference presentations. While practices likely vary by different organizations and for TIGs within AEA, the single case presented here is intended to promote thinking about best practices and encourage future research and discussion on the topic, specifically within the evaluation community. This paper also adds to the larger peer review literature by illustrating a case in which peer review was systematically applied for conference purposes.

The case grew out of discussions between graduate students and new evaluators during annual AEA conferences. Recognizing the importance of selecting the best contributions for the annual conference, the TIG leadership sought to improve the old system as it was believed inadequate to serving this purpose. After elaborating on the need for a revision of the formerly used process, the article discusses results of the 2006 review process, and concludes with implications for TIG review processes, AEA, and the field in general.

The 2006 Graduate Student & New Evaluator TIG Review Process

The GS&NE TIG was initiated in 1999 with the purpose to develop activities for and to represent the interests of graduate students and those new to the field of evaluation. Specifically, the TIG intended to promote communication, networking, and other means to increase opportunities for new professionals to engage in the field of evaluation. After the 2005 AEA conference, the GS&NE TIG Board began debating the relative merit of the current proposal review process, suggesting a need for revising the formerly used review process.

The Need for Revision

The discussion about the validity of the existing review process primarily involved the TIG board and to a lesser extent other members of the TIG. First and foremost was the board's desire to increase the quality of accepted proposals, and thereby, conference presentations. AEA conference participation, TIG business meetings, and other events provided feedback about the perceived quality of presentations. For example, graduate students stated that they submit proposals to the GS&NE TIG if they felt the quality is insufficient for acceptance in other, more prestigious TIGs. Additionally, TIG members reported attending sessions that were visited by less than a handful of people. Moreover, the annual AEA conference evaluations documented concerns with the quality of conference presentations (Barnett, Costantino, Hood, Jang, & Walker, 2004; Bartholomay et al., 2001; Mason et al., 2004; Swindler et al., 2002).

Second, and although many formats for reviewing proposals exist across AEA TIGs (Schröter, November, 2007), the proposals submitted to the GS&NE TIG were originally not blinded and were only reviewed by TIG board members. Given that proposers of submissions, and respective authors, chairs, and discussants were often known to one or more of the reviewers, this was especially problematic. The TIG's board and membership understood that this poorly designed single-blind review process made it more likely that proposals by well-known personae or reviewer peers would be accepted.

Finally, the TIG board felt that it was difficult to make decisions based on the AEA review criteria.[1] While the use of the standard form is not mandatory, AEA encourages its use to serve as the foundation of the review processes. Like many others, the GS&NE TIG found that although some of the criteria were useful, the form as a whole did not accurately capture the TIG's values and intended purpose (Barnett, Costantino, Hood, Jang, & Walker, 2004; Bartholomay et al., 2001; Mason et al., 2004; Swindler et al., 2002).

Proposal Review Process Revisions

In 2006, the board began searching for alternative review processes and ultimately commenced the development of a review method that would be as unbiased and reliable as possible. To establish an impartial review process, the TIG board decided to remove names and affiliations of all involved in a given submission. Moreover, reviewers of submissions were to be selected systematically, so that a reviewer would not have the same affiliation as contributors on any given submission. This process minimized recognition of peers and other evaluation personae under review (Ross et al. 2006). While some would argue that the work of peers can be recognized otherwise, the 2007 review process suggested that only 2% of reviewers thought they knew the submitting proposer or institution.

Additionally, the GS&NE board revised the AEA review sheet to include criteria that more accurately reflected the TIGs' purpose and values as well as quality and relevance, aspects inherent in the general and specific logic of

evaluation. Among others, a literature review was conducted and relevant stakeholder perspectives (i.e., graduate students and new evaluators) were incorporated to ensure inclusiveness of the process. Moreover, the criteria were designed to strike most evaluators as highly plausible - that is, characterizing a proposal of high quality. Finally, many of these criteria emerged from the second author's dissertation research, which, in part, set forth to identify and verify some of the properties that characterize good, valuable, and important research of various types and classes (Coryn, 2007).

The New Review Instrument

The newly constructed review instrument consisted of ten items that were assessed on a five-point scale to provide information that would discriminate proposals of higher quality and relevance from those with lower quality or relevance, and yield a ranking across all proposals. The TIG opted for a five-point scale in order to permit reviewers to select the "golden middle" or "undecidedness" for any particular criterion.

This section was followed by three decision items. The first asked for holistic recommendations as to whether the submission should be accepted or rejected. Second, if reviewers recommended a submission for the conference, they were asked to indicate the extent to which the proposal reflected the 2006 conference theme. The third item asked if the proposal should be nominated to the presidential strand.

Finally, the review instrument included two open-ended questions designed to obtain reviewers' rationales for these decisions and to receive feedback and additional comments.

Comparison to AEA's Review Sheet

The revised GS&NE TIG review instrument contained some criteria which were consistent with the AEA form as well as modifications and additions. The criteria present on both the AEA and newly developed TIG review instruments included *relevance/importance to a broad AEA audience*, *relevance/importance to the TIG*, and *innovativeness*. Unanimously, the TIG agreed that an AEA presentation should have relevance to the GS&NE TIG specifically, and the AEA audience at large, to maximize attendance at TIG-sponsored sessions. Innovativeness for the purpose of the GS&NE TIG peer review process meant that the proposal included something new.

The seven remaining criteria on the new TIG review worksheet comprised variations and/or addenda to AEA's standard form. Variations included items on the significance of the submission to evaluation (a) as a discipline, (b) practice, (c) theory/logic, and (d) methodology. This is in contrast to the AEA review sheet, which focuses on how evaluation methods, theories, policies, and practices are evaluated on a scale from "very focused on findings" to "very focused on practice." This scale was not useful for our purposes for a variety of reasons. First, the compartmentalization into four criteria assumes that different foci cannot co-occur. Evaluation research, however, often falls into both categories. Almost all of the studies on evaluation have been related to or focused on practice (e.g., Alkin & Daillak, 1979; Chandler & Henderson, 2001; Christie 2003, 2007; Christie & Masyn, 2007; Cousins & Leithwood, 1986; Patton et al., 1977; Preskill & Caracelli, 1997; Preskill, Zuckerman, & Matthews, 2003; Rockwell, Dickey, & Jasa, 1990; Shaddish & Epstein, 1987; Thompson, Brown, & Furguson, 1981; Williams, 1989). However, the intent of most submissions related to evaluation research is focused on presenting study findings. In these situations, this scale offers no mechanism for valuing the quality or significance of proposals. That is, the extent to which it impacts evaluation as a discipline and practice. Evaluation research, again, helps illuminate the

problematic nature of this scenario. Although prior research focused on various aspects of evaluation, recent evaluation research has been highly focused on evaluation practice (e.g., Azzam, 2007; Barela, 2006; Christie, 2003, 2007; Christie & Masyn, 2007).

Another variation from the standard AEA review sheet was the adaptation of the criterion *technical quality*, which is defined by AEA as "a proposal that meets high standards of technical quality as defined by the TIG. Expanded to a manuscript, a very high quality proposal would likely be published in a peer-reviewed journal." The TIG felt that although not all valuable submissions lend themselves for expansion into manuscripts, this criterion reflected an "overall judgment" about proposal quality that might reflect publication potentials. As such, the TIG included an item, called *overall judgment* in which ratings on all other criteria had to be taken into account. This does not mean that other good submissions are not worthy of consideration for the conference, but that quality is an important factor in deciding which presentation will generate interest in attendees.

Two criteria were added to the review sheet: *originality* and *creativity*. Originality was described as "the proposal reflects independent thought or constructive imagination." Arguably, originality is related to innovativeness. However, the criterion has been added to emphasize a different manifestation, namely independent though that is beyond an innovative twist on something that is already practiced within the evaluation community. As Guetzkow, Lamont, and Mallard (2004) note, originality is defined narrower in the literature on the sociology of science (natural sciences) than in the social sciences and humanities. However, because evaluation is a transdiscipline (Coryn & Hattie, 2006), both definitions must be considered. The broad definition comprises innovativeness. That is, the proposer must not necessarily present new thought to the discipline, but to the TIG, and more specifically, the respective reviewer. For example, a presentation might focus on testing a theory of evaluation. The theory may not be new, the method used for testing may not be new, and the findings may have been illuminated elsewhere, but the session is still viewed as an innovative contribution to the conference as it can confirm the existent knowledge base. In contrast, originality as used within the TIG review sheet is meant to embrace the narrow definition were the proposal under review must illustrate that something unique is being presented. Our review of the 2006 AEA criteria sheet indicated that this distinction had not been made and that reviewers generally struggled with the difference. However, both were included within the analysis to emphasize (i.e., weight) their importance.

Creativity was also added to the sheet and was described as "the proposal reflects unconventional means." Again, this criterion has similarities with the two previous ones. However, creativity can be reflected in aspects of a proposal that are unconventional, for example, in terms of their structure and approach to the proposed presentation (e.g., approaches to organizing panels, roundtables, and think tanks, but also in unconventional means for proposing papers and posters). Moreover, creativity is also part of the notion of research as an autonomous pursuit, free of interference by sponsors or other interested stakeholders, and sometimes presents itself as research or evaluation outside of conventional paradigms. Finally, it can be argued that both research and evaluation are creative endeavors as much as they are scientific, and it would be self-defeating to attempt to constrain or fail to recognize creativity (Bush, 1945, 1960).

The criterion *diversity* was omitted from the standard form within the revised TIG review sheet. In assessing "diversity," AEA's standard sheet describes the construct as "a proposal's contribution to the diversity of presentations with respect to subject matter, populations, programs, methods, culture, ethnicity, and presenters." The rationale for excluding this

rather important item is three-fold. First, a reviewer who is assigned a set of two to three submissions is unlikely to be able to judge the degree of diversity (i.e., heterogeneity or variation) of submissions received by the TIG in general, or the conference as a whole, in terms of subject matter, populations, programs, methods, or culture. Second, reviewers who assess blinded submissions are unable to judge diversity in terms of presenter race/ethnicity and experience. Finally, as submissions are reviewed within a specific TIG, submissions should reflect the interests of the group in some way, thus sharing commonalities with the TIG as defined by the TIG's values and motives.

The 2006 Review

In 2006, the GS&NE TIG received 26 proposal abstracts for review, which comprised 2.4% of the total submissions sent to AEA. The majority of submissions were papers, followed by posters, workshops, and other sessions such as debates, demonstrations, multi-papers, roundtables, and think tanks (see Table 1). Because of the TIGs purpose, lectures and panels are not usually submitted to the TIG. A majority of our submissions fall into one of two categories: proposals submitted by graduate students or novice evaluators and proposals that address thematic issues (e.g., advice for those seeking careers in evaluation) pertinent to our stakeholders.

As is the case with other TIG program chairs, prior to the review process, the incoming program chair participated in telephone training provided by AEA to generate an understanding of the general process and associated tasks for the individual TIGs. As a result of the training, the GS&NE TIG program chair recruited volunteers via the TIG's listserv. At the commencement of the 2006 AEA review process, the TIG program chair received a document including the proposals, a list of AEA members who volunteered for proposal review, a list of volunteers to chair sessions, and the AEA review form. All 32 volunteers, including those from the list provided by AEA and the recruited ones were invited to participate in the TIG's review process.

Table 1
Submissions to 2006 AEA Conference and GS&NE TIG by Category

Category	AEA	GS&NE	% GS&NE
Debate	2	1	50.0%
Demonstration	71	1	1.4%
Lecture	18	—	0.0%
Multi-paper	63	1	1.6%
Panel	160	—	0.0%
Paper	532	13	2.4%
Poster	91	6	6.6%
Roundtable	94	1	1.1%
Workshop	21	2	9.5%
Think Tank	35	1	2.9%
Total	**1,087**	**26**	**2.4%**

The Process. Prior to sending out submissions, the lead program chair removed all key identifying information (e.g., names and affiliation) from each proposal and randomly assigned two to three proposals to each volunteering reviewer. Random assignment took place after controlling for institutional independence (i.e., a reviewer assigned to a submission from his/her respective institution). The lead TIG board members (i.e., chair, chair-

elect, program chair, and program co-chair) were asked to evaluate all abstracts, excluding those from their own institution or known personally. After assignments were made, each reviewer received an e-mail invitation, including the submissions and the revised proposal review

sheet. Reviewers were asked to provide feedback within a three week period. In the event that they could not meet the deadline, reviewers were asked to reply immediately, so that the submissions could be reassigned. Four volunteer reviewers were unable to participate and their tasks were reassigned. As a result, 28 reviewers from 20 different institutions participated in the process and each of the 26 proposal abstracts was evaluated by five to seven reviewers (174 reviews total).

Evaluation and ranking of proposals. Completed reviews were entered into a database. A five step procedure was used to evaluate and rank the reviews and either accept, downgrade, or reject proposals. First, as shown in Table 2, a simple rate of acceptance per proposal was calculated as a percentage, where *percentage of acceptance* (column 5 in Table 2) was

$$= \frac{n\,\text{accept}}{n\,\text{accept} + n\,\text{reject}} \tag{1}$$

Second, the arithmetic mean for each of the ten Likert-type items was calculated for each proposal based on the number of proposal reviewers, simply as

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \tag{2}$$

Third, the ten item averages for each item per proposal were summed to produce an *item average total score* (column 6 in Table 2) calculated as

$$= \bar{x}_1 + \bar{x}_2 + \cdots \bar{x}_{10} \tag{3}$$

Thus, these scores had a potential range from 0 to 50.

Fourth, the percentage of the total possible score was determined by dividing the total score by the total possible score (i.e., 50) for each proposal, producing an *item percentage score* (column 7 in Table 2), where this score was

$$= \frac{\text{item average total score}}{50} \tag{4}$$

Fifth, the arithmetic mean of each proposal's *percentage of acceptance* and *item percentage score* was calculated (i.e., the average of columns 5 and 7) to produce a *combined percentage score* (column 8 in Table 2). The rationale underlying this analytic procedure was to make use of the full range of information provided by the reviews in the decision making process.

Table 2
Summary of GS&NE TIG Ranking Procedure by Proposal

Proposal	Number of Reviewers	Accept	Reject	Percentage of Acceptance	Item Average Total Score	Item Percentage Score	Combined Percentage Score	Decision
1	7	1	6	14%	29.85	60%	37%	Reject
2	7	3	4	43%	32.90	66%	54%	Reject
3	6	3	3	50%	31.00	62%	56%	Reject
4	6	3	3	50%	33.00	66%	58%	Reject
5	7	4	3	57%	30.63	61%	59%	Reject
6	7	4	3	57%	33.40	67%	62%	Reject
7	7	4	3	57%	33.90	68%	62%	Reject
8	7	5	2	71%	34.20	68%	70%	Downgrade
9	6	5	1	83%	32.00	64%	74%	Downgrade
10	6	5	1	83%	32.50	65%	74%	Downgrade
11	5	4	1	80%	34.17	68%	74%	Downgrade
12	7	6	1	86%	31.80	64%	75%	Downgrade
13	7	6	1	86%	33.20	66%	76%	Downgrade
14	7	6	1	86%	34.50	69%	77%	Downgrade
15	7	6	1	86%	34.60	69%	77%	Downgrade
16	7	6	1	86%	37.60	75%	80%	Downgrade
17	6	6	0	100%	33.20	66%	83%	Accept
18	7	7	0	100%	35.00	70%	85%	Accept
19	7	7	0	100%	35.80	72%	86%	Accept
20	7	7	0	100%	36.25	73%	86%	Accept
21	7	7	0	100%	36.80	74%	87%	Accept
22	7	7	0	100%	37.20	74%	87%	Accept
23	6	6	0	100%	37.75	76%	88%	Accept
24	6	6	0	100%	38.20	76%	88%	Accept
25	7	7	0	100%	38.40	77%	88%	Accept
26	5	5	0	100%	42.50	85%	93%	Accept

Analysis of internal consistency of the 10 Likert-type items produced a Cronbach's α = .94. As shown in Equation 5, interrater reliability, based on raters' accept or reject decisions (columns 3 and 4 in Table 2) was .80, and was estimated as a *coefficient of agreement* represented by the total proportion of observations (P_o) of which there was agreement, or

$$P_o = \frac{\text{number of exact agreements}}{\text{number of possible agreements}} = \frac{\sum f_o}{N} \qquad (5a)$$

$$= \frac{136}{171} = .80 \qquad (5b)$$

where $\sum f_o$ is the sum of the frequencies of observed agreements, and N is the number of pairs of scores obtained.

Finally, all proposals were sorted by rank order of combined percentage score and a decision was rendered using the rubric described below (column 8 in Table 2).

Decision Rubric. To summarize the findings, ensure consistency in the review process, and yield credible, valid decisions, a rubric was developed collaboratively by the GS&NE TIG board members. The resulting standards to be applied to reviewed proposals were the following:

> *Reject*. Proposals with a combined percentage score < 69%. (Note that these proposals were rejected by at least three reviewers.)
>
> *Downgrade*. Proposals with a combined percentage score ≥ 70% and ≤ 80%. (Note that these proposals were rejected by at least one but no more than two reviewers.) If these proposals had a format that was not to be downgraded (e.g., a poster), the submission was kept.
>
> *Accept*. Proposals with a combined percentage score > 80%. (Note that these proposals were not rejected at all.)

While the cut scores appeared high, the rational for the standards applied was to assure fairness. In essence, we felt that if we were to decline one proposal with three reviewer rejections, all others in this category should be rejected as well. Similarly, we proceeded with downgrading and accepting submissions.

Results. The above review process yielded the following results: 27% of abstracts were rejected, 35% were downgraded, and 38% were accepted. These numbers were somewhat higher than averages reported across all AEA TIGs in 2001 (Bartholomay et al., 2001) and for the GS&NE TIG in previous years. Bartholomay and colleagues (2001), for example, found that on average fewer than 10% of proposals are rejected across all AEA TIGs, with a range from less than 10% up to 30%. In general, AEA's suggested rejection rate is a function of conference location and varies from seven to 15% (personal communication with Susan Kistler, June 29, 2007). A study of peer review for the American Heart Association's annual Scientific Sessions (Ross et al., 2006) suggest that these numbers are much lower than those of other scientific communities and their respective meetings.

Implications

The present article has a number of potential implications for TIG review processes, AEA, and the field in general. We believe that the presented method lends *credibility* to the proposal review process. As suggested by Scriven more than a decade ago (1991), "Evaluations often need to be not only valid but such that their audience will believe that they are valid…This may require extra care about avoiding (apparent) conflict of interest…" (pp. 110-111). The method presented here explicitly attempts to avoid such conflicts of interest by ensuring, through a blind peer-review process and systematic reviewer selection, that only presentations of the highest quality are selected for the conference. Moreover, it included a fairness principle in its rubric. As such, the mechanism increases the trustworthiness of the peer review within the TIG, while also giving due consideration to other ethical concerns.

The process also warrants *equity*. All individuals, regardless of rank, past contribution to the field, gender, ethnicity, culture, or affiliation, have an equal opportunity of presenting if their proposal reflects substantial merit, worth, and significance to a TIG and the broader AEA audience. The revision of the review sheet, while not perfect by any means, is a first step toward best practice in the evaluation of proposals for the annual AEA conference. To fully realize the potential requires that the evaluation community continues to put their best efforts forward, particularly those who have previously molded the discourse in some way.

Blinding proposals ensures that prospective presentations submitted by TIG board members, their colleagues, or recognized evaluation experts are not automatically accepted. From experience, the authors know that a mediocre proposals submitted by an important name or affiliation is hard to assess in a credible way. This happens at the cost of rejecting submissions which may have been of higher quality, but written by a no-name author.

This is in some respects similar to review panels in the United States awarding up to 20 additional points for random assignment experimental designs (the so-called "20% solution") over well-designed non-experimental projects when reviewing proposals for funding (Coryn, Hattie, Scriven, & Hattie, 2007; Julnes & Rog, 2007). In short, the process presented in this paper creates less of an opportunity for the conference and field to be perceived as a "good ol' boys (and girls) club" (Benos et al., 2007; Ross, et al., 2000). While AEA is known for its inclusionary efforts (thus, in some respects AEA's criterion of diversity), credibility and transparency are necessary features to promote the health of the profession.

This is not a trivial matter, such that accepted proposals impact discussions and thinking on and about evaluation, as well as overall conference quality. The initial analyses of a forthcoming study that examines AEA TIG review processes in 2007 found that more than two-thirds of AEA's TIGs do not engage in such practices (Schröter, 2007). However, if all TIGs would use similar, and ideally the same, procedures, TIG-level findings could be aggregated to AEA-level. Then, conference sessions could be selected by relative quality of TIG proposals, rather than by the size of the TIG. While AEA suggest that the number of proposals to be accepted is a function of TIG size and size of conference facilities, a process that clearly differentiates between higher quality and lower quality submissions may yield a conference that only presents the best work in the field, thus reflecting practices of a truly evaluative organization. Of course, the diversity criterion must be reintroduced on the top level to allow new and evolving TIGs to emerge. Some would argue that such practices are especially harsh for graduate students and new evaluators, but as the discussions in the GS&NE TIG have shown, these individuals want to present good work and be treated equitably.

Assuming that many attendees participate in AEA's annual conference to pursue professional development and networking opportunities, it is important that high quality is reflected throughout to allow best practice to flourish in the varying sectors. Furthermore, although an increased rejection rate might impact the conference size, it also allows for greater transparency for selecting sessions from the conference program and minimizes the occurrence of presenters delivering long-planned contributions to less than a hand-full of individuals who actually perceive the presentation as worthy of attending. As such, the revised process impacts worth.

Additionally, this method, or a refined version of it, has the potential to propel the AEA Board and TIGs to continually refine their scope and purpose. As contexts evolve, so too do fields, programs, and people. A more formalized process might assist TIG and AEA Board members in thinking about and defining important topics. Perhaps more importantly, this process has the potential to impact evaluation theory, methodology, practice, and research, by including originality, innovativeness, and creativity as criteria that call for unique perspectives or contributions (regardless of their size or impact). Assessing innovativeness, originality, and creativity as somewhat overlapping but different concepts puts greater weight on the overall construct and ensures that conference quality is held to similar standards as that of academic journals. As a result, increased numbers of papers and presentations that take place at the annual conference may be published in our journals.

Of course, not all presenters desire to publish their work, but the review process may yield increased audience interest and feedback that could stimulate the author to continued work toward publication. Surely, not everything that sparks interest and growth in the evaluation community must be published, but it may impact how we as evaluators improve our own work by exposing it to others.

Although this paper offers an alternate method for reviewing proposals, it also raises some questions. First, is the presented method easily implemented across all TIGs or are further revisions necessary? While we believe that the described review process is an improvement to the old system used by the GS&NE TIG, it too can still be improved. As implemented in 2006, the process was very time-consuming on the side of TIG Board members. Streamlining the process might require the development of an online system that reduces the time for data entry and analysis, for example. To alleviate potential concerns with the method and elucidate best practice in AEA proposal evaluations, the GS&NE TIG is currently conducting a study that investigates TIG practices throughout AEA (Schröter, 2007). Furthermore, a think tank at the upcoming AEA conference is planned to illuminate current thought of TIGs and the larger AEA membership. This think tank may also inform the second question, that is, what is the purpose of the annual conference and corresponding review processes for TIGs and AEA? If the purpose of the conference is to promote dialogue about evaluation theory, methodology, and practice; and to present thinking and research on and about evaluation, then, all involved should begin to assess our own performance on these domains. A side effect of promoting the presentation of the best submission may be increased networking capability between peers; newcomers and oldtimers; and theoreticians, methodologists, and practitioners. In doing so and if we find room for improvement, we believe this method is one of many that have the potential to assist us in better meeting current and future goals.

Note

1. The review criteria were presented on a rating sheet and consisted of six items to be rated on varying four-point Likert-type scales. Briefly, the items include: (a) relevance/importance to a broad AEA audience, (b) relevance/importance to the TIG, (c) technical quality, (d) innovativeness, (e) diversity in terms of subject matter, populations, programs, methods, culture, ethnicity, and presenters, and (f) focus on evaluation methods, theories, policies, and practices. For those interested, a copy of the review form discussed in this paper can be obtained by contacting the first author at daniela.schroeter@wmich.edu.

Author Acknowledgments

We would like to thank the numerous individuals who provided input for revising the review sheet as well as those who volunteered in the actual review process.

References

Aksnes, D. W. (2005). *Citations and their use as indicators in science policy: Studies of validity and applicability issues with a particular focus on highly-cited papers*. Unpublished doctoral dissertation, University of Twente, Netherlands.

Alkin, M. C., Daillak, R. H. (1979). A study of evaluation utilization. *Educational Evaluation and Policy Analysis, 1*(4), 41-49.

Azzam, T. (2007). *Evaluator contextual responsiveness*. Unpublished doctoral dissertation, University of California, Los Angeles.

Barela, E. (2005, October). *How school district evaluators make sense of their practice: A folk theory*. Paper presented at the Joint Conference of the Canadian Evaluation Society and the American Evaluation Association, Toronto, Canada.

Barnett, E., Costantino, T., Hood, L., Jang, E. E., & Walker, K. C. (2004). *Evaluation of the 2003 annual conference of the American Evaluation Association: Final report*. Retrieved on May 10, 2007 from http: //www.eval. org/training/eval history.asp

Bartholomay, T., Lilligren, L., Smith, J., Volkov, B., Williems, G., & King, J. A. (2001). *AEA 2001 Conference evaluation: Final report*. Retrieved on May 10, 2007 from http:// www.eval.org/training/evalhistory.asp.

Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., LaFrance, M., et al. (2007). The ups and downs of peer review. *Advances in Physiological Education, 31*. 145-152.

Burnham, J. C. (1992). How journal editors came to develop and critique peer review procedures. In H. F. Mayland & R. E. Sojka (Eds.), *Research ethics, manuscript review and journal quality* (pp. 55-62). Madison, WI: ACS Miscellaneous Publication.

Bush, V. (1960). *Science—The endless frontier: A report to the President on a program for postwar scientific research*. Washington, DC: National Science Foundation. (Original work published in 1946).

Campanario, J. M. (1998). Peer review for journals as it stands today—Part 1. *Science Communication*, *19*(3), 181-211.

Ceci, S. J., & Peters, D. P. (1982). Peer review—a study of reliability. *Change*, *14*, 44–48.

Chandler, M. (2001, November). *How evaluators engage theory and philosophy in their practice*. Paper presented at the annual conference of the American Evaluation Association, St. Louis, MO.

Chelimsky, E. (1997). The political environment of evaluation and what it means for the development of the field. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 53-68). Thousand Oaks, CA: Sage.

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), *New directions for evaluation: The practice-theory relationship in evaluation* (Vol. 97, pp. 7-35). San Francisco, CA: Jossey-Bass.

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions. *American Journal of Evaluation, 28*(1)*,* 8-25.

Christie, C. A., & Masyn, K. E. (2007). *Latent profiles of evaluators' self-reported practices.* Manuscript submitted for publication.

Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. *Science*, *214*, 881-886.

Coryn, C. L. S. (2006). The use and abuse of citations as indicators of research quality. *Journal of MultiDisciplinary Evaluation, 3*(4), 115-120.

Coryn, C. L. S. (2007). *Evaluation of researchers and their research: Toward making the implicit explicit*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Coryn, C. L. S., & Hattie, J. A. (2006). The transdisciplinary model of evaluation. *Journal of MultiDisciplinary Evaluation*, *3*(4), 107-114.

Coryn, C. L. S., Hattie, J. A., Scriven, M., & Hartmann, D. J. (2007). Models and mechanisms for evaluating government-funded research: An international comparison. *American Journal of Evaluation*, *28*(4), 437-457.

Coryn, C. L. S., & Scriven, M. (2008). *Reforming the evaluation of research. New Directions for Evaluation:* San Francisco, CA: Jossey-Bass.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56*(3), 331-364.

Einstein, A. (1905a). Über einen die Erzeugung and Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik*, *17*(2), 132-148.

Einstein, A. (1905b). *Eine neue Bestimmung der Moleküldimensionen*. Unpublished doctoral dissertation, Zürich Universität, Zürich.

Einstein, A. (1905c). Über die von der molekülarkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, *17*(2), 549-560.

Einstein, A. (1905d). Zur Electrodynamik bewegter Körper. *Annalen der Physik*, *17*(2), 891-921.

Einstein, A. (1905e). Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig? *Annalen der Physik*, *18*(2), 639-641.

Foltz, F. A. (2000). The ups and downs of peer review: Making funding choices for science. *Bulletin for Science, Technology & Society*, *20*(6), 427-440.

Garcia, R., & Calantone, R. (2002). A critical look at technological innovation typology and innovativeness terminology: A literature review. *The Journal of Product Innovation Management*, *19*(2), 110-132.

Guetzkow J., Lamont M., & Mallard G. (2004). What is originality in the humanities and the social sciences? *American Sociological Review, 69*(2), 90-212.

Hackett, E. J. (1997). Peer review in science and science policy. In M. S. Frankel & J. Cave (Eds.), *Evaluating science and scientists: An East-West dialogue on research evaluation in post-Communist Europe* (pp. 51-60). Budapest, Hungary: Central European University Press.

Julnes, G., & Rog, D. L. (Eds.). (2007). Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation. *New Directions for Evaluation*, 113. San Francisco, CA: Jossey-Bass.

Justice, A. C., Cho, M. K., Winker, M. A., Berlin, J. A., & Rennie, D. (1998). Does masking author identity improve peer review quality? A randomized controlled trial. *JAMA*, *280*(3), 240-242.

Langfeldt, L. (2002). *Decision-making in expert panels evaluating research: Constraints, processes and bias* (Doctoral dissertation, University of Oslo, Norway). ISBN 82-7218-465-6.

Mainguy, G., Motamedi, M. R., & Mietchen, D. (2005). Peer review—The newcomer's perspective. *PLoS Biology*, *3*(9), 1534-1535.

Mason, G., Blanton, S., McDonald, K., Neal, J., Tanyu, M., Taylor-Ritzler, T., & Reeves, E. (2004). *Evaluation 2004: Overall conference survey*. Retrieved on May 10, 2007 from http://www.eval.org/training/evalhistory.asp.

McCook, A. (2006). Is peer review broken? *The Scientist*, *20*(2), 26.

McEneaney, J. E. (2001). *Electronic submission and review of NRC conference proposals: Re-engineering social literacies for online environments*. Paper presented at the meeting of the National Reading Conference. December 7, 2001. San Antonio, TX. Retrieved on June 10, 2007 from ttp://personalwebs.oakland.edu/~mceneane/nrc/conf01/NRCePropos als.doc.

Niederhauser, D.S., Wetzel, K., & Lindstrom, D. L. (2004). From manuscript to article: Publishing educational technology research. *Contemporary Issues in Technology and Teacher Education,4*(2), 89 -136.

OECD (1987). *Evaluation of research: A selection of current practices*. Paris, France: Organisation for Economic Co-Operation and Development.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q., Grimes, P. S, Guthrie, K. M, Brennan, N. J., French, B. D, & Blyth, D. A. (1977). In search of impact: An analysis of the utilization of federal health evaluation research. In C. H. Weiss (Ed.), *Using social research in public policy making* (pp. 141-164). Lexington, MA: D.C. Heath.

Preskill, H., & Caracelli, V. (1997). Current and developing conceptions of use: Evaluation

use TIG survey results. *American Journal of Evaluation,18*(1)*,* 209-225.

Preskill, H., Zuckerman, B., & Matthews, B. (2003). An exploratory study of process use: Findings and implications for future research. *American Journal of Evaluation, 24*(4)*,* 423-442.

Pyenson, L., & Sheets-Pyenson, S. (1999). *Servants of nature: A history of scientific institutions, enterprises, and sensibilities*. New York, NY: W. W. Norton & Company.

Rockwell, S. K., Dickey, E. C., & Jasa, P. J. (1990). The personal factor in evaluation use: A case study of a steering committee's use of a conservation tillage survey. *Evaluation and Program Planning, 13*(4)*,* 389-394.

Ross, J. S., Gross, C. P., Desai, M. M., Hong, Y., Grant, A. O., Daniels, S. R., Hachinski, V. C., Gibbons, R. J., Gardner, T. J., & Krumholz, H. M. (2006). Effects of blinded peer review on abstract acceptance. *JAMA*, *295*(14), 1675-1680.

Schröter, D. C. (2007). *Learning from AEA TIG proposal review standards*. Think tank at the annual conference of the American Evaluation Association, Baltimore, MD.

Scriven, M. (1991). *The evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

Scriven, M. (1993). Hard-won lessons in program evaluation. *New Directions for Program Evaluation*, *58*. San Francisco, CA: Jossey-Bass.

Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice*, *17*(2), 151-161.

Scriven, M. (2003). Evaluation in the new millennium: The transdisciplinary vision. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the millennium* (pp. 19-42). Mahwah, NJ: Lawrence Erlbaum Associates.

Shadish, W. R., & Epstein, R. (1987). Patterns of program evaluation practice among members of the Evaluation Research Society and Evaluation Network. *Evaluation Review, 11*(5)*,* 555-590.

Smith, R. (1999). Opening up BMJ peer review: A beginning that should lead to complete transparency. *BMJ*, *318*, 4-5.

Swindler, S., Hughes, G., Briggs, C., Yamazaki, K., Ohse, D., Pinero, S., & Sagrestano, L. (2002). *American Evaluation Association annual conference evaluation: Evaluation 2002 final report*. Retrieved on May 10, 2007 from http://www.eval.org/training/evalhistory.asp.

Turney, J. (1990). End of the peer show? *New Scientist*, *22*, 38-42.

Walsh, E., Rooney, M., Appleby, L., & Wilkinson, G. (2000). Open peer review: a randomized controlled trial. *The British Journal of Psychiatry*, *176*, 47-51.

Williams, J. E. (1989). A numerically developed taxonomy of evaluation theory and practice. *Evaluation Review, 13*(1)*,* 18-31.