Journal of MultiDisciplinary Evaluation
Volume 10, Issue 23, 2014

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Evaluating Application of Knowledge and Skills: The Use of Consensus Expert Review to Assess Conference Abstracts of Field Epidemiology Training Participants

Boris Volkov
Centers for Disease Control and Prevention and Oak Ridge Institute for Science and Education

Goldie MacDonald
Centers for Disease Control and Prevention

Dionisio Herrera
Training Programs in Epidemiology and Public Health Interventions Network

Donna Jones
Centers for Disease Control and Prevention

Mahomed Patel
Australian National University

Background: Often evaluations of training programs are limited — with many focusing on the aspects that are easy to measure (e.g., reaction of trainees) without addressing the important outcomes of training, such as how trainees applied their new knowledge, skills, and attitudes. Numerous evaluations fail to measure training's effect on job performance because few effective methods are available to do so. Particularly difficult is the problem of evaluating multisite training programs that vary considerably in structure and implementation from one site to another.

Purpose: NA

Setting: NA

Intervention: NA

Research Design: We devised a method of a consensus expert review to evaluate the quality of conference abstracts submitted by participants in Field Epidemiology Training Programs – an approach that can provide useful information on how well trainees apply knowledge and skills gained in training, complementing data obtained from other sources and methods. This method is practical, minimally intrusive, and resource-efficient, and it may prove useful for evaluation practice in diverse fields that require training.

Data Collection and Analysis: NA

Findings: NA

Keywords: *evaluation of training; multisite evaluation; consensus expert review; abstract quality*

Introduction

Training programs are often evaluated to demonstrate value to decision makers and other stakeholders and to improve training implementation and outcomes. However, evaluation seldom includes an assessment of how well trainees apply their newly acquired knowledge and skills. To quote Wargnier's (2012, p. 2) lament, evaluation methods these days are "systematically used to measure the satisfaction of participants, often used to measure knowledge acquisition, rarely used to measure changes in professional behaviors, and almost never used to measure the impact on business performance."

One reason evaluators fail to measure training's effect on job performance is that few effective methods or tools are available to do so. Particularly difficult is the problem of evaluating multisite training programs that vary considerably in structure and implementation from one site to another. Multisite evaluations need an efficient method that 1) can evaluate at least one activity common to all training sites; and 2) provides comparable data on how well trainees are likely to apply what they have learned when they return to their usual work. Recently, evaluators of the Field Epidemiology Training Program (FETP) devised a method to address these challenges. The purpose of this manuscript is to describe the use of a consensus expert review to evaluate the quality of conference abstracts submitted by participants in Field Epidemiology Training Programs in 10 countries – an approach that may be a useful addition to methods of evaluating similar training programs, since it produces information on how well trainees apply knowledge and skills gained in training. This method is practical, minimally intrusive, and resource-efficient, and it may prove useful for evaluation practice in diverse fields that require training. We hope that this article contributes to evaluation methodology and interests a broad audience of practitioners and researchers in the fields of evaluation and training.

Multisite Evaluation of Field Epidemiology Training Programs

The Centers for Disease Control and Prevention (CDC) in the United States began assisting other countries to develop epidemiology training programs in 1975 and modeled those training programs on CDC's Epidemic Intelligence Service (EIS). The purpose is to increase the number and quality of public health workers worldwide. A typical Field Epidemiology Training Program is a 2-year applied epidemiology training-through-service program. By providing trainees not only with classroom instruction, but also with field experience in responding to disease outbreaks, natural disasters, and other public health priorities, FETPs provide public health service while also building a workforce of trained epidemiologists (CDC, 2013; Patel & Phillips, 2009).

An FETP is a competency-based training, with at least 75% of the participants' (also called *fellows* or *residents*) time devoted to applying their new knowledge or skills in the field under the supervision of an expert field epidemiologist. Even while in training, FETP participants build or increase the public health capacity of their host country (CDC, 2006).

Although several FETPs have been evaluated (e.g., Bhatnagar et al., 2012; Patel, 2011) and the importance and value of FETPs are well-documented (Music & Schultz 1990; López & Cáceres, 2008; Schneider et al., 2011; Traicoff et al., 2008), experts do not agree on the best way to evaluate the quality of these programs: the terminology, frameworks, and indicators used for key program areas are inconsistent, and academic literature on options for productive evaluation of these programs is scarce.

Conducted in collaboration between CDC and the Training Programs in Epidemiology and Public Health Interventions Network (TEPHINET), the 2012-13 Multisite Evaluation of Field Epidemiology Training Programs was the first systematic study in more than 10 years that looked in a standardized and structured way at FETP implementation and proximal outcomes across multiple sites (Jones et al., 2014). A diverse group of program stakeholders determined the following purposes of the evaluation: to document key aspects of program design and implementation and to demonstrate accountability for use of resources and results. The evaluation used multiple methods and data sources. For example, original data were collected during the visits to the FETPs, through in-person interviews with FETP trainees and graduates, resident advisors, FETP directors, and other stakeholders and review of local documents. The evaluation combined these original data sources with secondary data sources, including documents on FETPs' development, planning, and implementation (as part of the written records available at CDC in Atlanta).

To assess the quality of participants' work, the evaluation team also developed and implemented

a blinded, systematic, consensus expert review of abstracts submitted to the 10th Global TEPHINET Conference (an FETP-specific conference). The FETP evaluators considered that the overall quality of the abstracts produced by an FETP's trainees was one of the indicators of the overall quality of the FETP training itself. One reason for using the quality of abstracts as an indicator was that the implementation of FETPs varies considerably from one FETP to another. For purposes of the evaluation, we needed at least one common activity across all FETP sites. When planning the evaluation, we learned that all FETPs required trainees to submit an abstract to the Global TEPHINET Conference and that almost all trainees did so.

The utility of our approach was also supported by the fact that writing an abstract to communicate important findings and messages about an applied research study at a scientific conference is an essential competency in field epidemiology. FETPs include training in conducting studies of public health-related events, threats, and challenges to generate an evidence base for informing decisions, policies, and public health actions. Trainees must therefore be effective at both carrying out epidemiologic studies and communicating their process and results. More specifically, quality of abstracts is considered a marker of competency in most of the 10 domains of core competencies relevant to FETPs: epidemiologic methods, biostatistics, public health surveillance, communication, prevention effectiveness, and epidemiology of priority diseases and injuries (CDC, 2006).

Furthermore, the abstract review method allowed us to move beyond merely assessing what trainees had learned (Level 2: *Learning* in Kirkpatrick & Kirkpatrick, 2010) to assessing changes in their behavior as a result of what they had learned (Level 3: *Behavior*) (i.e., we could assess how they applied the new knowledge or skills gained by participating in the FETP). Thus evaluating abstracts, in combination with the other methods of evaluation, would contribute to a better understanding of each FETP's overall training quality.

Using Abstract Quality Review in the Multisite Evaluation

For the multisite evaluation of FETPs, the "quality of abstracts" indicator was operationalized as "scientific rigor and merit of abstracts submitted by fellows or residents to the 2010 Sixth Global TEPHINET Conference (as determined by a panel of experts)" (Jones et al., 2014). The abstract quality was to be measured by consensus scores and ratings given by the panel of experts that reviewed a sample of the conference abstracts.

In collaboration with TEPHINET, the evaluation team selected three subject matter experts and invited them to be on the abstract review panel. The experts' qualifications included extensive epidemiology experience working in global settings; strong FETP experience; applied and academic research experience; and extensive experience reviewing conference abstracts. Each reviewer was located in a different geographic region and participated in the review via e-mail and telephone.

For this multisite evaluation, the evaluation team selected 10 FETPs that represented a broad spectrum of experience: FETPs from low- and middle-income countries, national and regional programs, long-standing and recently established programs, and programs with university affiliations and programs without such affiliations. For a program to be selected, it needed 1) to have a CDC-supported resident advisor on site and 2) to have graduated at least two cohorts of trainees. The FETPs participating in the evaluation had a wide geographic representation, with three regional and seven national FETPs, situated in Africa, Asia, Central America, and Eastern Europe.

Several factors influenced the decision to select the 2010 TEPHINET conference as a source of abstracts. First, all FETPs participating in the evaluation are part of the TEPHINET's international network and take part in this biennial conference. The TEPHINET's scientific conferences are important to FETPs' host countries and partner organizations: they combine scientific sessions and workshops related to managing public health systems and training programs, and FETP trainees benefit from the experience of presenting their work to an international audience of experts. Second, TEPHINET was an implementing partner for the multisite evaluation and, as a sponsor of the conference, could readily provide access to the conference abstract database. Third, the FETP cohorts participating in the evaluation had submitted their abstracts to this particular conference.

A member of the evaluation team (not a member of the review panel) selected a random sample of abstracts from those that the 10 evaluation sites had submitted to the 2010 Global TEPHINET conference. Total sample size was 49 abstracts (5 abstracts from each of nine FETPs and four from one FETP). To ensure an unbiased process for reviewing all abstracts, authors' names

and all geographic identifiers were removed from each abstract prior to review, and the abstracts were provided randomly to the reviewers so that abstracts from the same FETP were not clustered.

Consensus among these independent experts was paramount to producing a credible determination of an abstract's quality. Studies of reviews indicate a serious problem of low levels of reviewer agreement (Cohen & Patel, 2006; Landis & Koch, 1977; Ragone et al., 2011; Rowe et al., 2006). Callaham & Tercier (2007) and Ragone et al. (2011) note that often there is a high degree of randomness in the review processes. In contrast, by using a consensus panel review approach, we sought to reduce the level of subjectivity, randomness, and resulting unreliability in the peer review processes. In contrast to simple averaging of individual reviewers' scores, consensus panel review requires the panelists to come to agreement, which means that reviewers must compromise to resolve conflicting perceptions of an abstract's quality.

We define the consensus expert review of abstracts as a structured analytic technique by which scientific abstracts can be rigorously and collaboratively scored and rated by an expert panel via standardized analysis, assessment, and comparison, with the ultimate goal of achieving a consensus rating for the abstracts. On the basis of lessons learned and suggestions found in published studies, our evaluation used the following guidelines for the abstract review process:

- A consensus-based approach.
- Highly experienced, diverse, and independent reviewers.
- Clear objectives and criteria for evaluating each abstract.
- A sensible scoring system with interval scales.
- An assessment focused exclusively on abstracts' quality.

The reviewers were asked to evaluate each abstract and then produce a summary assessment of the overall quality of the abstracts submitted by each FETP. The reviewers created and pilot-tested a point/scoring system, which was based on previous systems used by CDC's EIS and by TEPHINET for reviewing conference abstract submissions. The review criteria were as follows:

1. Rationale for conducting the study and the study objectives.
2. Methods: were they appropriate for addressing the study objectives, and did they conform to requirements for conducting a sound scientific study?
3. Results: were they appropriate and valid, and did they address the study objectives?
4. Conclusions: were they related appropriately to the results?
5. Public health significance of the work described.
6. Usefulness of the study and the effect or potential effect of the findings and recommendations.
7. Overall clarity of the abstract.

Each criterion was evaluated on a scale of zero to four, with a maximum score of 28. The reviewers used the seven review criteria to answer three review questions:

1. *Did the authors do the right thing?* That is, did the authors give adequate and relevant reasons for conducting their study? This question is related to the applied component of field epidemiology; it is about the relevance and actual or potential usefulness of the study and its contribution to the field of public health or epidemiology (assessed on the basis of criteria 1, 5, and 6).
2. *Did the authors do it the right way?* That is, did the authors take the right steps to answer their research question? This question is related to the scientific merits of the study and complements the applied component of field epidemiology referred to in Question 1 (assessed on the basis of criteria 2, 3, and 4).
3. *Is the writing clear, and does the text follow a logical sequence?* This question refers to the communication skills of the authors. Clarity and logic are important because it is possible that, even if the authors did the right thing in the right way, they lack the skills to explain the study's methods and findings effectively to the reader (Criterion 7).

The scores for each of the seven review criteria and for the overall abstract score were categorized into four groups (based on a maximum score of 4 for each of the 7 review criteria). Review criteria scores were categorized as: 0 = poor, 1 = weak, 2 = fair, 3 = good, 4 = very good. Total scores for the overall abstract were categorized as: 0–7 = poor, 8–14 = fair, 15–21= good, 22–28 = very good. Each reviewer assigned a score to each abstract

and commented on the quality relating to each of the seven criteria, as well as comments on how the abstract could be improved, after which all reviewers discussed their ratings, resolved any disagreements, and came to consensus about the final score for the abstract. To reach consensus on the total score awarded to each abstract, they scanned the total score given for each abstract by the three reviewers to assess the concordance between the scores. A variation of up to 2 points between the scores was accepted to reflect a 'consensus' score; where the total score given by a reviewer differed by 3 or more points from any other reviewer, the abstract was reassessed by all reviewers to decide whether the score could be modified to reach consensus.

Subsequently, the mean, median, and distribution of scores and ratings were determined for all abstracts from each FETP. Using this combined information, reviewers assigned each FETP's overall composite rating on abstract quality as *very good, good, fair,* or *poor*. The reviewers documented in an Excel worksheet the original total scores, the revised total (or consensus scores), the final scores for each review area given by each reviewer for each abstract, and each FETP's overall ratings of abstract quality.

Use of the Review Findings to Inform Program Planning

The expert review determined considerable variation in quality of the conference abstracts within and across the sample of FETPs. Table 1 shows the data on abstract quality by FETP site, providing the range of scores for each FETP's individual abstracts and a median score for each FETP (FETPs are identified only by letter codes). The overall range of individual scores was 6-24 (out of a possible 28), whereas median scores by FETP ranged 9-20. The table also shows the overall composite ratings of quality of abstracts *(very good, good, fair, poor)* for each FETP: such ratings for five FETPs (P2, P3, P4, P7, P10) fell in the *good* or *very good* range, while the other five programs' quality of abstracts (P1, P5, P6, P8, P9) was *fair* or *poor*.

Table 1
Quality of Abstract Scores and Ratings by FETP: Multisite Evaluation of FETPs, June 2012–February 2013

FETP	RANGE OF SCORES	MEDIAN SCORE	OVERALL RATING
A	15-22	20	VERY GOOD
B	17-21	19	VERY GOOD
C	12-24	18	GOOD
D	9-21	15	GOOD
E	10-23	12	GOOD
F	10-19	14	FAIR
G	11-17	13	FAIR
H	11-16	13	FAIR
I	8-17	11	FAIR
J	6-20	9	POOR

Our evaluation also looked at how the findings from the abstract review related to the data on selected characteristics of the 10 programs. For example, FETPs with high abstract quality ratings also had high levels of "resident advisor engagement" (i.e., training-related interaction between trainees and the FETP's resident advisor from CDC), a critical component of the training process and trainees' development of requisite competencies.

The process and findings of the expert review of abstract quality were relevant to program stakeholders, and the findings were an important component of the report of results for the overall multisite evaluation. They provided a measure of the quality of one aspect of the programs and participants' core work that helped to inform the discussions by decision makers and other stakeholders about ongoing development, planning, and implementation of the FETP model of training.

Another value of the abstract review was the review panel's recommendations to TEPHINET about changes to the guidelines for preparing and

scoring abstracts for future TEPHINET conferences. As a result, a team of FETP experts prepared a more detailed guide to help FETP staff and trainees work through the steps of writing and reviewing abstracts to be submitted for presentations at scientific conferences. The new guide was based on the strengths, weaknesses, and other lessons learned during the consensus expert review of abstracts.

Conclusion and Implications for Evaluation Practice

Often evaluations of training programs or activities are limited or short-sighted — with many focusing their evaluation on the aspects that are easy to measure (e.g., reaction of trainees) without addressing the important outcomes of training, such as how trainees applied their new knowledge, skills, attitudes, and sense of confidence from the training. A consensus expert review of conference abstracts can usefully address at least two of the 12 checkpoints of training identified by Scriven's (2010) *The evaluation of training: A checklist approach:* "learning" and "application." According to Scriven, for the checkpoint "learning," we "need evidence that participants in fact mastered (at least much of) the intended content, and acquired the intended value or attitude modifications" (p. 8). The "application" checkpoint is intended to find "whether participants appropriately used, and continued to use appropriately, what they learned from the training in their work context," which may involve, among other factors, an "examination of work product" of a trainee (p. 10). Our experience shows that a rigorous, structured, review of scientific abstracts can provide an indication of the effect the training will have on participants' professional knowledge and work and how they will apply their newly acquired skill and knowledge.

Abstract reviews can provide useful evaluation information that complements data obtained from other sources and methods. At the same time, the method is "resource-efficient" in terms of costs related to the experts' travel, time, and data collection. The panel members can review abstracts from their home or any other location via telephone, e-mail, or Internet-based communication (e.g., Skype). The method allows evaluators to use secondary data, which may significantly reduce the cost and time normally required for acquiring original data. Gaining access to conference abstracts does not seem complicated; abstracts can be obtained from conference organizers or training staff. Selecting a random sample of abstracts as we did for this multisite evaluation may significantly decrease the time and resources required for evaluating a training program while maintaining high quality for the evaluation.

One limitation to using this method is in that abstract quality was determined on the basis of a pool of abstracts submitted by trainees, but we had no way of knowing the level to which the FETP's resident advisor reviewed the abstract or was involved in writing it. To overcome this challenge and draw the most accurate picture of the status of the program, we interpreted the findings for the abstract quality indicator with caution and triangulated the findings on abstract quality with findings from other sources and for other FETP indicators.

Obviously, the abstract reviews for training evaluation purposes will be more useful and valid when the conference organizers have clear, detailed guidelines and standards for writing and submitting abstracts and associated papers. Nevertheless, even in the absence of such guidelines, a review is helpful in judging the overall scientific rigor, common sense, and logic of the abstracts.

Especially when resources for evaluation are limited, this method is a practical, minimally intrusive, and relatively low-cost alternative to other kinds of assessment that are impractical because of inaccessible or unreliable data, prohibitive costs, or ethical issues. Perhaps such abstract reviews could also be used as a kind of "screening test" or "early warning sign" of the status of a training program's functioning and quality or as a method of ongoing evaluation of the progress of a program. Properly customized, reviewing abstracts may prove useful for strengthening evaluation practices in diverse fields that require training, in addition to field epidemiology and public health.

Our experiences with using this approach and the resulting benefits gained seem to provide strong support for its face validity. However, more research is needed to evaluate and validate the effectiveness of abstract reviews in different evaluation contexts. Such research could also seek to explain the convergence or divergence of the results for the abstract quality, other quality indicators, and the overall training quality.

References

Callaham, M., & Tercier, J. (2007). The relationship of previous training and experience of journal peer reviewers to subsequent review quality. *PLOS Medicine*, *4*(1): e40.

Centers for Disease Control and Prevention. (2006). *Field Epidemiology Training Program Development Handbook*. Atlanta, Georgia: Author. Retrieved from: http://www.cdc.gov/globalhealth/fetp/pdf/fetp_development_handbook_508.pdf

Centers for Disease Control and Prevention. (2013). *FETP - History and Value of Establishing Field Epidemiology Training Programs*. Retrieved from: http://www.cdc.gov/globalhealth/fetp/history.html

Jones, D., MacDonald, G., Volkov, B., & Herrera-Guibert, D. (2014). Multisite Evaluation of Field Epidemiology Training Programs: Findings and Recommendations. Centers for Disease Control and Prevention: Atlanta, Georgia. Retrieved from: http://www.cdc.gov/globalhealth/fetp/pdf/fetp_evaluation_report_may_2014.pdf

Cohen, I., & Patel, K. (2006). Peer review interrater concordance of scientific abstracts: A study of anesthesiology subspecialty and component societies. *Anesthesia & Analgesia*, *102*(5), 1501-1503.

International Program for Development Evaluation Training (IPDET). (2007). *IPDET Handbook. Module 7. Approaches to Development Evaluation. 2007.* Retrieved from: http://www.worldbank.org/ieg/ipdet/modules/M_07-na.pdf

Kirkpatrick, J., & Kirkpatrick, W. (2010). *Training on Trial*. New York, NY: AMACOM.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

López, A., & Cáceres, V. (2008). Central America Field Epidemiology Training Program (CA FETP): A pathway to sustainable public health capacity development. *Human Resources for Health*, 6:27.

Music, S., & Schultz, M. (1990). Field epidemiology training programs: New international health resources. *Journal of American Medical Association, 263*, 3309-3311.

Patel, M. (2011). *Vietnam's Field Epidemiology Training Program: An evaluation of the program for the first class of fellows 2009-2011*. Unpublished report.

Patel, M., & Phillips C. (2009). Strengthening field-based training in low and middle-income countries to build public health capacity: Lessons from Australia's Master of Applied Epidemiology Program. *Australia and New Zealand Health Policy*, 6:5.

Ragone, A., Mirylenka, K., Casati, F., & Marchese, M. (2011). A quantitative analysis of peer review. In *Proceedings of the 13th conference of the international society for scientometrics and informetrics (issi)*. Durban: issi.

Rowe, B., Strome, T., Spooner, C., Blitz, S., Grafstein, E., & Worster, A. (2006). Reviewer agreement trends from four years of electronic submissions of conference abstract. *BMC Medical Research Methodology*, 6:14.

Schneider, D., Evering-Watley, M., Walke, H., & Bloland, P. (2011). Training the global public health workforce through applied epidemiology training programs: CDC's experience, 1951-2011. *Public Health Reviews. 33*, 190-203.

Scriven, M. (2010). *The evaluation of training: A checklist approach*. Retrieved from: http://michaelscriven.info/images/EVALUATION_of_TRAINING.11-27-11.2.pdf

Traicoff, D., Walke, H., Jones, D., Gogstad, E., Imtiaz, R., & White, M. (2008). Replicating success: developing a standard FETP curriculum. *Public Health Reports, 123*, S28-34.

Wargnier, J. (2012). *Evaluating and demonstrating the value of training: Challenges, practices and trends at the age of new learning technologies*. Retrieved from: http://www.CrossKnowledge.com