

American Journal of Evaluation

<http://aje.sagepub.com>

Toward a Taxonomy of Essential Evaluator Competencies

Jean A. King, Laurie Stevahn, Gail Ghere and Jane Minnema

American Journal of Evaluation 2001; 22; 229

DOI: 10.1177/109821400102200206

The online version of this article can be found at:

<http://aje.sagepub.com/cgi/content/abstract/22/2/229>

Published by:

 SAGE Publications

<http://www.sagepublications.com>

On behalf of:

American Evaluation Association

Additional services and information for *American Journal of Evaluation* can be found at:

Email Alerts: <http://aje.sagepub.com/cgi/alerts>

Subscriptions: <http://aje.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations (this article cites 11 articles hosted on the
SAGE Journals Online and HighWire Press platforms):

<http://aje.sagepub.com/cgi/content/refs/22/2/229>

Forum contributions present essays, opinions, and professional judgments. Forum articles speak to and about the philosophical, ethical, and practical dilemmas of our profession. By design the Forum is open to diverse views, in the hope that such diversity will enhance professional dialogue. Standard citations and reference lists should be used to acknowledge and identify earlier contributions and viewpoints. Manuscripts should typically not exceed 15 double-spaced typewritten pages in length unless the paper is invited by the Editor.

Toward a Taxonomy of Essential Evaluator Competencies

**JEAN A. KING, LAURIE STEVAHN, GAIL GHERE, AND
JANE MINNEMA**

ABSTRACT

This article discusses an exploratory study designed to determine the extent to which evaluation professionals, representing diverse backgrounds and approaches, could reach agreement on a proposed taxonomy of essential evaluator competencies. Participants were 31 diverse individuals in the field of program evaluation in the greater Minneapolis-St. Paul, Minnesota area who systematically engaged in a Multi-Attribute Consensus Reaching process. Both quantitative and qualitative results predominantly indicated consensus on more than three-fourths of the proposed competencies. Areas of disagreement reflected the role- and context-specific nature of evaluation practice.

INTRODUCTION

Over the past two decades, a series of interrelated activities have propelled program evaluation toward maturity as a profession. Worthen and Sanders (1991) identify five historical trends that have shaped this professionalization: (1) the emergence of career

Jean A. King • 330 Wulling Hall, 86 Pleasant St. SE, University of Minnesota, Minneapolis, MN 55455; Tel: (612) 626-1614; Fax: (612) 624-3377; E-mail: kingx004@umn.edu. The authors were equal contributors both to the study and to the writing of this article.

American Journal of Evaluation, Vol. 22, No. 2, 2001, pp. 229–247. All rights of reproduction in any form reserved.
ISSN: 1098-2140 Copyright © 2001 by American Evaluation Association.

opportunities in evaluation, (2) the development of preparation programs for evaluators, (3) the institutionalization of evaluation in American education, (4) the development of evaluation as distinct from other professions, and (5) methodological developments. In addition, Worthen and Sanders (1991) identify “touchstones” that mark evaluation as a distinct profession, including the adoption of *The Program Evaluation Standards* (Joint Committee on Standards for Educational Evaluation, 1994). Despite these developments, however, several evaluators (Merwin & Weiner, 1985; Patton, 1990; Worthen, 1999; Worthen & Sanders, 1991) continue to question whether the field of program evaluation can yet claim full professional status.

One noticeable deficiency is a coherent and widely accepted set of the unique skills and knowledge that distinguishes professional evaluators. For example, although *The Program Evaluation Standards* constitute a comprehensive set of standards that provide parameters for high-quality evaluation processes and products, the standards do not directly address the competencies an evaluator needs to function effectively in specific contexts. Smith (1999) notes that the *Guiding Principles of the American Evaluation Association* (AEA) (American Evaluation Association, 1995) are also of little use in this regard because there is no way to derive specific skills and knowledge from such overarching principles. A key question is whether or not the evaluation community can reach agreement on a set of evaluator competencies given the diverse philosophical and practical approaches that exist within the field (Smith, 1999; Worthen, 1999). Indeed, as the evaluation profession has grown and matured over the past 20 years, it has envisioned its own Holy Grail: a set of essential competencies that evaluators in diverse settings employing diverse methods would agree are essential to their practice.

In the context of evaluator certification, Smith (1999, p. 525) identifies three possibilities for developing evaluator competencies: (1) asking evaluation professionals to generate a list, (2) conducting job analyses of different tasks and activities that users or employers who receive evaluation products consider crucial, or (3) using AEA’s *Guiding Principles* to identify competencies. She then details the limitations of each of these approaches—the professionals’ list may only include a limited number of competencies and may unfairly limit competition from evaluators who did not participate in developing the list; job analyses are costly; and, as noted, the *Guiding Principles* are simply too general to be of use (Smith, 1999, pp. 525–526).

Nonetheless, several frameworks identifying evaluator tasks and skills have been proposed (e.g., see Anderson & Ball, 1978; Covert, 1992; Mertens, 1994; Sanders, 1979; Worthen, 1975). Content also has been proposed for preparatory and staff development programs in evaluation (e.g., see Altschuld, 1995; Ingle & Klaus, 1980; Sanders, 1986). None of the tasks, skills, or content frameworks, however, have been derived from a systematic process or validated by empirical consensus building among diverse professionals in the field. These initial frameworks, along with proposals for evaluator certification, have sparked healthy debates at conferences and in the evaluation literature about the value and feasibility of developing a comprehensive set of evaluator competencies (e.g., see Altschuld, 1999; Altschuld & Bickman, 1998; Jones & Worthen, 1999; Smith, 1998, 1999; Worthen, 1999), but discussions to date have fallen short of reaching consensus. Worthen (1999, p. 547) summarizes conventional wisdom about the search for competencies: “. . . [I]t would seem far too early to predict whether our field’s conceptions will soon reach the point where there will be a cohesive core of competencies that can be widely agreed to . . . Only time will tell whether such a development is even a responsible hope.”

The purpose of this article is to discuss an exploratory study designed to assess the extent to which evaluation professionals who came from diverse backgrounds and approaches reached agreement on a proposed taxonomy of competencies, which we call the *Essential Evaluator Competencies* (King, Minnema, Ghere, & Stevahn, 1999). To determine agreement on the perceived importance of the competencies, participants engaged in a systematic consensus-reaching procedure whose fundamental guiding question was: What are the salient knowledge, skills, and attitudes necessary to function as an effective evaluator?

We envisioned the study as exploratory and therefore began by recruiting individuals active in the field of program evaluation in the greater Minneapolis-St. Paul, Minnesota, area who represented a range of diversity on such evaluator characteristics as formal evaluation training, experience as a practicing evaluator, evaluator roles and responsibilities, type of organization, and so on (described in the next section). We reasoned that starting with diverse evaluators primarily in one state would provide valuable preliminary information for determining whether to pursue systematic inquiry in broader arenas. Although the presence of some shared attributes may make consensus within any given state more feasible than across larger regions, consensus among diverse evaluators within a state is surely not guaranteed, if we are to believe the literature. Furthermore, lack of consensus within a state almost certainly would make consensus across larger areas unlikely.

DETERMINING FACE VALIDITY

Developing the *Essential Evaluator Competencies*

The *Essential Evaluator Competencies* are a comprehensive set of proposed evaluator competencies that began innocently enough as a class exercise in an evaluation studies colloquium. Advanced graduate students engaged in an inductive thinking activity in which they identified essential evaluator competencies by using a concept formation strategy (described in Joyce & Weil, 1996). The consensus evident in that class, which contained evaluators from a number of diverse settings, inspired four of those present (the professor and three advanced doctoral students in evaluation studies and educational psychology at the University of Minnesota) to develop the competency list more systematically (see King et al., 1998, 1999). The taxonomy's further development took place in two phases.

The first phase involved additional work on the initial list of critical evaluator competencies. Taking the list developed in class, we began by reviewing (1) the literature on evaluator competencies, (2) *The Program Evaluation Standards* (Joint Committee on Standards for Educational Evaluation, 1994), and (3) the *Guiding Principles of the American Evaluation Association* (American Evaluation Association, 1995). We then created a draft version of the *Essential Evaluator Competencies* (which comprised the knowledge, skills, and attitudes proposed to be essential to function as an effective program evaluator) and obtained feedback on the draft from two respected evaluation experts, one of whom is a nationally recognized leader in the field.

The second phase of further development involved piloting and revising the draft version of the *Essential Evaluator Competencies*. We conducted two separate pilots in which a total of six practicing professional evaluators and two advanced evaluation studies students engaged in a Multi-Attribute Consensus Reaching (MARC) process (described below) to

determine the degree of consensus on the perceived importance of the competencies in the taxonomy. After each pilot we revised the taxonomy, adding new competencies that the participants suggested and rearranging or rewording others based on their feedback. The second revision became the *Essential Evaluator Competencies* used in this study. The taxonomy (shown in Table 1) identifies 4 broad competency *domains* (designated by Roman numerals), 16 competency *categories* (designated by capital letters within the domains), and 49 competency *items* (designated by Arabic numerals within the categories).

Establishing Face Validity

We assumed that diverse individuals involved in the field of program evaluation would have valid opinions about what constitute essential competencies for the practice of evaluation and that they were, therefore, the right group to establish the face validity of the proposed *Essential Evaluator Competencies*. We chose to use a MACR process, a variation of the Multi-Attribute Consensus Building (MACB) process designed to facilitate group decision making, to engage participants in discussing the competencies (for a complete description of MACB, see Vanderwood & Erickson, 1994).

In our MACR process, participants in small group sessions of 3 to 10 individuals were given the *Essential Evaluator Competencies*. First, participants individually judged the perceived importance of the competencies using a Likert-type rating system (described in the next section). Second, the ratings for each competency were entered into a computer spreadsheet that calculated means and ranges. High means and narrow ranges indicate greater agreement on the perceived importance of the competency in question; because of the MACR anchor requirement, explained in the next section, narrower ranges will always contain high means. In contrast, wide ranges may contain either low or high means, depending on how the frequency distribution clusters. Regardless of the mean, a wide range indicates disagreement on the perceived importance of the competency. Third, participants discussed the calculated results and articulated varying rationales underlying their ratings. The discussions were designed to allow each competency to get a "fair hearing" so that participants would consider alternative and sometimes opposing viewpoints before making final judgments on the perceived importance of each competency. Finally, participants individually provided their final ratings for each competency, making any desired changes from their previous ratings. The ultimate results provided both quantitative and qualitative data on agreement/disagreement on the perceived importance of the competencies, thereby providing an empirical measure of face validity for inclusion of the competencies in the taxonomy. After combining the data from the various sessions, we applied decision rules (described in the Results section of this article) and used reasoned judgment to distinguish what constituted "real agreement" versus "real disagreement." More agreement indicated greater face validity of the given competency, whereas more disagreement indicated less face validity.

The Rating System

The MACR process requires competencies to be numerically rated within their respective clusters (see Table 1). Accordingly, participants first rate the perceived importance of the *domain cluster* composed of competencies I, II, III, IV. Next participants rate the *category clusters* within each domain (i.e., the cluster of IA, IB, IC within the first domain; IIA, IIB, IIC within the second domain; IIIA, IIIB, IIIC, IIID, IIIE within the third domain; and IVA,

TABLE 1.
Essential Evaluator Competencies: Means and Ranges

<i>Competencies</i>	<i>Domains (Roman numerals)</i>		<i>Categories (Capital letters)</i>		<i>Items (Arabic numerals)</i>	
	<i>Mean</i>	<i>Range</i>	<i>Mean</i>	<i>Range</i>	<i>Mean</i>	<i>Range</i>
I. Systematic Inquiry	95.10	60–100				
IA. Able to do research-oriented activities*			87.10	50–100		
IA1. Framing the research question(s)					94.03	10–100
IA2. Research design					90.23	50–100
IA3. Measurement					80.00	20–100
IA4. Research methods (quantitative, qualitative, and mixed methods)					92.65	70–100
IB. Able to do evaluation-oriented activities			97.26	70–100		
IB1. Evaluation theory, models, and underlying philosophical assumptions					86.61	0–100
IB2. Needs assessment					91.58	60–100
IB3. Framing the evaluation question(s)					99.97	99–100
IB4. Evaluation design					97.32	80–100
IB5. Evaluation processes					97.61	90–100
IB6. Making judgments*					74.68	20–100
IB7. Developing recommendations*					82.16	50–100
IB8. Meta-evaluation					78.06	10–100
IC. Able to do activities common to both research and evaluation			94.58	75–100		
IC1. Literature review*					80.58	10–100
IC2. Sampling					82.16	0–100
IC3. Instrument construction					94.90	50–100
IC4. Data collection					95.71	80–100
IC5. Data analysis					94.65	80–100
IC6. Data interpretation					97.90	80–100
IC7. Reporting results					96.45	80–100
II. Competent Evaluation Practice	94.35	55–100				
IIA. Able to serve the information needs of intended users			96.94	50–100		
IIB. Able to do situational analysis			95.48	75–100		
IIB1. Knowledgeable about organizational development, change, and politics					87.29	0–100
IIB2. Able to analyze the political context of an organization					93.87	80–100
IIB3. Respectful of the uniqueness of the evaluation site and client					91.94	50–100

(continued)

TABLE 1.
(Continued)

<i>Competencies</i>	<i>Domains</i>		<i>Categories</i>		<i>Items</i>	
	<i>(Roman numerals)</i>		<i>(Capital letters)</i>		<i>(Arabic numerals)</i>	
	<i>Mean</i>	<i>Range</i>	<i>Mean</i>	<i>Range</i>	<i>Mean</i>	<i>Range</i>
IIB4. Open to others' input					93.23	50–100
IIB5. Able to adapt/change study as needed					96.45	50–100
IIC. Able to organize and manage evaluation projects			98.06	80–100		
IIC1. Able to respond to a request for proposal					78.71	10–100
IIC2. Able to write formal agreements					84.65	0–100
IIC3. Able to budget an evaluation					87.58	0–100
IIC4. Able to access needed resources (e.g., information, personnel, instruments)					95.29	50–100
IIC5. Able to supervise others					79.42	0–100
IIC6. Able to train others					81.71	0–100
IIC7. Able to conduct the evaluation in a non-disruptive manner*					90.65	50–100
IIC8. Able to complete work in a timely manner					94.06	50–100
IIC9. Able to deal with stress during a project*					89.52	50–100
III. General Skills for Evaluation Practice	91.61	60–100				
IIIA. Logical and critical thinking skills			97.58	50–100		
IIIB. Written communication skills			92.90	60–100		
IIIC. Verbal communication skills			95.81	60–100		
IIID. Interpersonal competence			94.19	75–100		
IIID1. Negotiation skills					90.13	75–100
IIID2. Conflict resolution skills*					86.45	50–100
IIID3. Group facilitation skills					87.10	0–100
IIID4. Group processing skills					87.26	0–100
IIID5. Teamwork/Collaboration skills					96.61	75–100
IIID6. Cross-cultural skills*					90.32	50–100
IIIE. Computer application skills*			84.84	50–100		
IV. Evaluation Professionalism	88.39	60–100				
IVA. Knowledge of yourself as an evaluator*			89.45	50–100		
IVB. Ethical conduct			99.52	85–100		
IVB1. Ensures the honesty and integrity of the evaluation					98.87	85–100
IVB2. Is able to convey to potential clients your evaluation approach and skills					91.77	65–100

(continued)

TABLE 1.
(Continued)

Competencies	Domains (Roman numerals)		Categories (Capital letters)		Items (Arabic numerals)	
	Mean	Range	Mean	Range	Mean	Range
IVB3. Respects the security, dignity, and self-worth of the respondents, program, participants, clients, and other stakeholders					98.71	90–100
IVB4. Is responsible for contributing to the general and public welfare*					73.19	40–100
IVC. Knowledge of professional standards (e.g., Joint Committee Standards & AEA <i>Guiding Principles</i>)			78.55	0–100		
IVD. Application of professional standards			86.13	0–100		
IVE. Professional development			91.19	70–100		
IVE1. Is aware of needs for professional growth					92.42	50–100
IVE2. Reflects on practice*					93.23	50–100
IVE3. Networks*					80.81	40–100
IVE4. Updates personal knowledge in the evaluation field (e.g., workshops, conferences, journals)					89.68	0–100
IVE5. Updates knowledge in relevant content areas*					89.52	50–100
IVE6. Contributes to the knowledge base of evaluation*					60.84	0–100

Note. Bold and asterisk (*) indicate “real” disagreement on perceived importance; see text for explanation.

IVB, IVC, IVD, IVE within the fourth domain). Finally, participants rate the *item clusters* within each category (i.e., the cluster of IA1, IA2, IA3, IA4 within the first category of the first domain; IB1, IB2, IB3, IB4, IB5, IB6, IB7, IB8 within the second category of the first domain; and so on throughout the entire taxonomy). For consistency, all participants were instructed to use a Likert-type scale to rate the perceived importance of each evaluator competency within each cluster. The scale contains values from 0 to 100 with designated intervals of perceived importance from *very unimportant* (0–20) to *very important* (80–100). The MACR process also requires at least one competency within each cluster to be rated 100; all other competencies within the cluster can be rated 0 to 100. This rating requirement serves two purposes. First, it provides a numeric anchor for calculating the range of ratings for each competency, making range comparisons among competencies possible because each consistently ends at 100. Second, it provides a numeric anchor necessary for calculating the weighted perceived importance of each competency relative to the others in the cluster.

After participants individually rate each competency within a given cluster, the ratings are entered into a computer, calculated, and displayed. Figure 1 shows a display for a sample cluster of competencies. The sample shows the mean and range calculated for each compe-

Competency	Person A	Person B	Person C	Person D	Person E	Mean	Range MIN-MAX
X	50	95	100	90	100	87	50-100
Y	100	100	100	95	90	97	90-100
Z	75	100	100	100	95	94	75-100

Figure 1. Sample Computer Display for *Evaluator Competency Ratings*.

tency. Large ranges indicate more variability (i.e., greater disagreement on perceived importance), whereas small ranges indicate less variability (i.e., greater agreement on perceived importance). After participants review the display, the MACR facilitator leads the group in a discussion by inviting individuals to articulate the rationales underlying their ratings. Typically, the discussion begins by focusing on the competency that has the largest range and, therefore, the greatest disagreement regarding perceived importance. In the sample display shown in Fig. 1, the discussion would begin with competency X.

The weighted perceived importance of each competency within any given cluster also provides a measure of how essential participants judge each competency to be. Figure 2 shows a sample of weighted importance calculations. The sum of the weighted importance calculations of the competencies within each cluster always equals 1. It is calculated by first summing the means and dividing each respectively by the sum. When the weighted importance calculations of the competencies within a cluster are nearly equivalent (as in this example), we can conclude that all competencies are perceived as equally important. However, when the weighted importance calculations of the competencies within a cluster are not nearly equivalent (e.g., if the weighted importance of X was 0.18, Y was 0.42, and Z was 0.40), we can conclude that those competencies with smaller weighted importance calculations are not perceived to be as important as the competencies with larger weighted importance calculations. Because we calculated the weighted importance of the competencies after all of the MACR sessions had been conducted, participants did not consider the weighted importance information in their decision making during the respective MACR sessions.

Participants

Participants were 31 individuals (3 men and 28 women) involved in the field of program evaluation in the greater Minneapolis-St. Paul region. They were recruited to participate in

Competency	Weighted Importance
X	.31
Y	.35
Z	.34
TOTAL	1.00

Figure 2. Sample Weighted Importance Calculations for *Evaluator Competency Ratings*.

the process of establishing face validity on the proposed set of essential evaluator competencies because they (1) were involved in program evaluation in the state, (2) represented diversity on a number of evaluator characteristics (described below), and (3) were in the Minneapolis-St. Paul area when the study was conducted. Recruitment strategies included telephone and e-mail contacts by using lists of evaluators from various organizations that conduct evaluations in Minnesota, as well as face-to-face invitations to individuals in the evaluation community. We purposefully sought and successfully recruited participants who represented diversity across the following characteristics (see Fig. 3):

- Years as a practicing program evaluator,
- Type of organization in which evaluation work primarily takes place,
- Job title/position,
- Evaluator role relevant to the organization,
- Age category,
- Familiarity with *The Program Evaluation Standards* (Joint Committee on Standards for Educational Evaluation, 1994), and
- Training courses completed related to program evaluation.

Data Collection and Analysis

We conducted five MACR sessions, with 6 individuals in the first session, 8 in the second, 10 in the third, 4 in the fourth, and 3 in the fifth. Each session lasted approximately 2.5 hours during which time participants were given directions, quantitatively rated each evaluator competency within designated clusters, qualitatively discussed rationales for ratings, and then indicated any desired changes in previous quantitative ratings. We combined the final ratings from all five sessions to calculate the overall mean, range, and weighted importance for each competency within each cluster. We also content analyzed all of the qualitative discussion data across the sessions to determine common themes.

Quantitative data analysis began by establishing criteria for interpreting the means and ranges. Once established, we used the criteria to classify the proposed competencies in the set as either “perceived important” or “perceived unimportant” for functioning as an effective evaluator. Larger variability in agreement indicated that people disagreed about the perceived importance of a competency, whereas smaller variability in agreement indicated that people shared similar views about the perceived importance of a competency (whether or not it was perceived as more or less important). We analyzed all of the evaluator competencies within their respective clusters and classified each competency accordingly. We also calculated the weighted perceived importance of each competency within each cluster to determine the extent to which the weightings were relatively equal.

Qualitative data analysis entailed determining common themes, issues, concerns, and points of emphasis in the discussions across sessions. In three of the sessions we audiotaped and transcribed verbatim the participants’ discussions; in one of the sessions two independent observers wrote detailed and extensive notes on the participants’ comments during the discussion (a tape recorder was not used); and in one of the sessions we, unfortunately, did not obtain discussion data because of audiotape equipment failure. We began the analysis first by reading the three transcripts and the one set of notes entirely to grasp the flow of the discussion within the respective sessions. Next, we read the text within each set of transcripts and the notes that corresponded to the clusters of competencies specified in the *Essential*

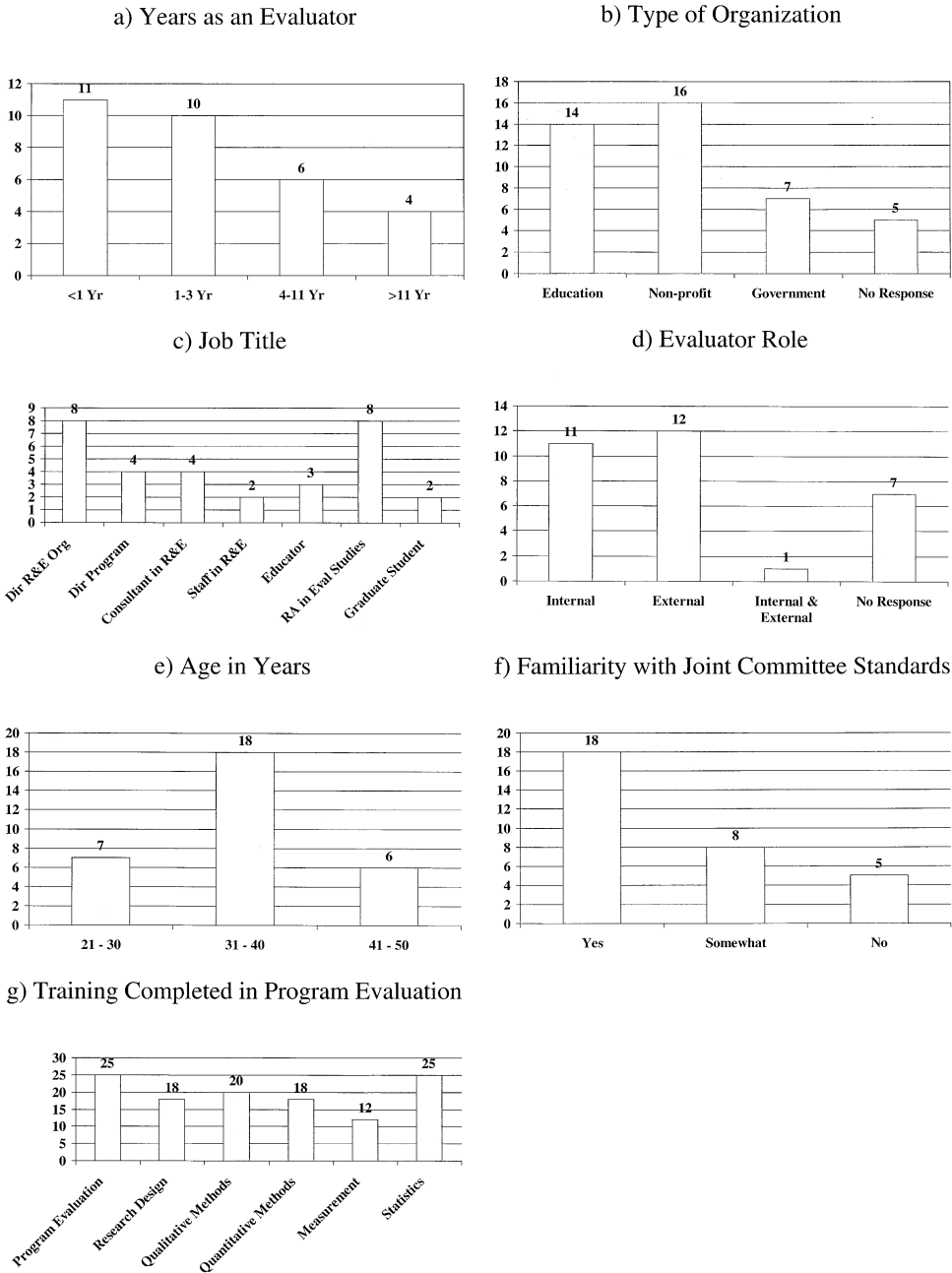


Figure 3. MARC Participant Characteristics.

Evaluator Competencies to examine in greater detail the common themes that emerged within each cluster. We conducted the qualitative analysis separately from the quantitative analysis, and then compared the results to determine the extent to which the qualitative data explicated the quantitative.

RESULTS

Quantitative

The ratings we were most interested in—disagreement on the perceived importance of competencies—were determined by applying the following criteria, identified after the pilot testing. First, competencies with means less than 80 were identified. Second, competencies with ranges larger than 60–100 were identified (i.e., the lowest value in the range was less than 60). Third, for competencies that met the first two criteria, frequency distributions were analyzed for outlier ratings. A given rating was identified as an outlier when (1) there was more than a 20-point gap between that rating and the next closest rating (on either side) in the frequency distribution, and (2) at least 90% or more of the cumulative percentage of ratings were 85 or higher in the frequency distribution. Competencies that met the first two criteria but had outlier ratings as specified by the third criterion were not considered to represent “real” disagreement on perceived importance (i.e., a single individual’s low rating did not result in a label of disagreement).

The results, presented in Table 1, distinguish between competencies on which there was “real” disagreement (indicated by bold and an asterisk) and those on which there was “real” agreement. A total of 15 competencies (3 *category* competencies and 12 *item* competencies) were classified as competencies with “real” disagreement. Table 2 presents sample quotations that illuminate issues related to these competencies, either in a single quotation or in two contrasting quotations. All other competencies in the *Essential Evaluator Competencies* (4 *domain* competencies, 13 *category* competencies, and 37 *item* competencies) were classified as competencies for which there was “real” agreement. In summary, there was “real” disagreement on the perceived importance of approximately 22% of the total competencies in the *Essential Evaluator Competencies* (15 out of 69) and “real” agreement on approximately 78% of the competencies in the taxonomy (54 out of 69).

It is noteworthy that there was nearly unanimous agreement on the perceived importance of two competencies. The first of these competencies was IB3, “Framing the evaluation question(s),” for which the mean was 99.97 with a range of 99–100. One participant’s comment summed up what participants consistently expressed across sessions: “If you don’t have a good evaluation question, nothing else matters.” The second of these competencies that participants valued strongly across all sessions was IVB, “Ethical conduct,” for which the mean was 99.52 with a range of 85–100. In one discussion, a participant said: “To me [ethics] are givens. Coming in [to this field] as a researcher, I remember this as intuitive.” In another discussion, when participants were invited to provide their rationales for uniformly giving ethics a rating of 100, an initial moment of silence was broken by one person’s comment that seemed to capture the sentiments of all: “Well, duh!”

The weighted importance of the competencies within each cluster was also calculated (see Fig. 4), suggesting the perceived importance that participants gave that competency compared to other competencies in its cluster. It is noteworthy that within nearly every

TABLE 2.
Sample Qualitative Responses: Issues with the *Essential Evaluator Competencies*

<i>Competencies</i>		<i>Comments</i>
IA	Able to do research-oriented activities	“I don’t want to confuse research and evaluation too much. We need to distinguish between those two, and when you are doing evaluation you are not necessarily doing research. You need to know the difference, but not necessarily know the research parts in detail.”
IB6	Making judgments	“When I started doing and reading evaluation, judgment was very important. The more I do, the less I think it is important and the less I want to do it. I don’t think an evaluator, an outsider, should make a judgment.”
IB7	Developing recommendations	“I think it should be defined [by] what the program wants. Who is commissioning the evaluation? If they say they want recommendations, then you give them recommendations. If they say they want judgments, then you give them judgments.”
IC1	Literature review	“I have given literature review 100. Before you can start, you should be aware of what others have done before you so you might take some of their instruments. I think it is very important.” “I rated [literature review] 78 because I think I thought it was kind of optional; helpful, but not necessary.”
IIC7	Able to conduct the evaluation in a non-disruptive manner	“Let me just clarify that the ‘non-disruptive’ is for the client?”
IIC9	Able to deal with stress during a project	“I have a question about the ability to deal with stress. That seems to be a given for everyone in every situation. But I’m wondering if ability to deal with ambiguity on a project is more to the point, perhaps?”
IIID2	Conflict resolution skills	“Conflict resolution between what? I guess I have a hard time putting it between evaluation team members, between yourself and stakeholders. It just seems too broad.”
IIID6	Cross-cultural skills	“I’m jumping in my seat. That’s such a hot button for me. I think there is way too much, well from what I’ve seen at the university, too much emphasis put on people having technical skills and not having the interpersonal. I think, particularly when you are working with community groups, that it’s really important to have good group facilitation skills and cross-cultural skills.” “I rated [cross-cultural skills] low just because I work in an organization that has not been diverse, unfortunately, and it just does not come up as often as it should.”

(continued)

TABLE 2.
(Continued)

<i>Competencies</i>		<i>Comments</i>
IIIE	Computer application skills	<p>“I think that is something you could subcontract out if you needed to. It’s not a necessity. You can also pick it up. Not like critical thinking skills or communication. You can pick up software, though.”</p> <p>“I rated computer skills a 70 as [you] can hire out for a computer application, such as a statistician, but you need the word processing.”</p>
IVB4	Responsible for contributing to the general public welfare	<p>“I think you could be a good evaluator without being responsible for the welfare of the world.”</p> <p>“I just see it as a Miss America, a wish for world peace. I don’t want that necessarily to be my responsibility. It might be my contribution, but I don’t think it is my responsibility.”</p>
IVE3	Networks	<p>“You can probably be a good evaluator and wouldn’t have to be out networking.”</p> <p>“It balances my hatred for responding to RFPs. By networking, I don’t have to respond to RFPs.”</p>
IVE6	Contributes to the knowledge base of evaluation	<p>“We’re not trying to change the world. We’re not in academia. Organizational culture we can influence, but the world of evaluation, not necessarily.”</p> <p>“I saw it as your experience in doing an evaluation could contribute to the knowledge base of evaluation and how it is conducted and what works and what doesn’t work.”</p>

cluster, the weighted importance of the competencies was approximately equal. There were two exceptions. Within the cluster of competency items under “Evaluation professionalism, Ethical conduct” (i.e., items IVB1, IVB2, IVB3, and IVB4), the weighted importance of item IVB4, “Is responsible for contributing to the general and public welfare,” was notably less than the weighted importance of the other items in that cluster. Similarly, within the cluster of competency items under “Evaluation professionalism, Professional development” (i.e., items IVE1, IVE2, IVE3, IVE4, IVE5, and IVE6), the weighted importance of item IVE6, “Contributes to the knowledge base of evaluation,” was notably less than the weighted importance of the other items in that cluster.

Qualitative

Two themes emerged from the content analysis of the qualitative data. The first theme was that participants across all sessions consistently made a broad distinction between evaluation and research. Although they acknowledged the role of research theory and methodology in program evaluation and appreciated how this knowledge can inform the evaluation process, they clearly did not view *research* as synonymous with *evaluation*. They repeatedly emphasized that their role was to evaluate, not to do research. One participant said: “An evaluator

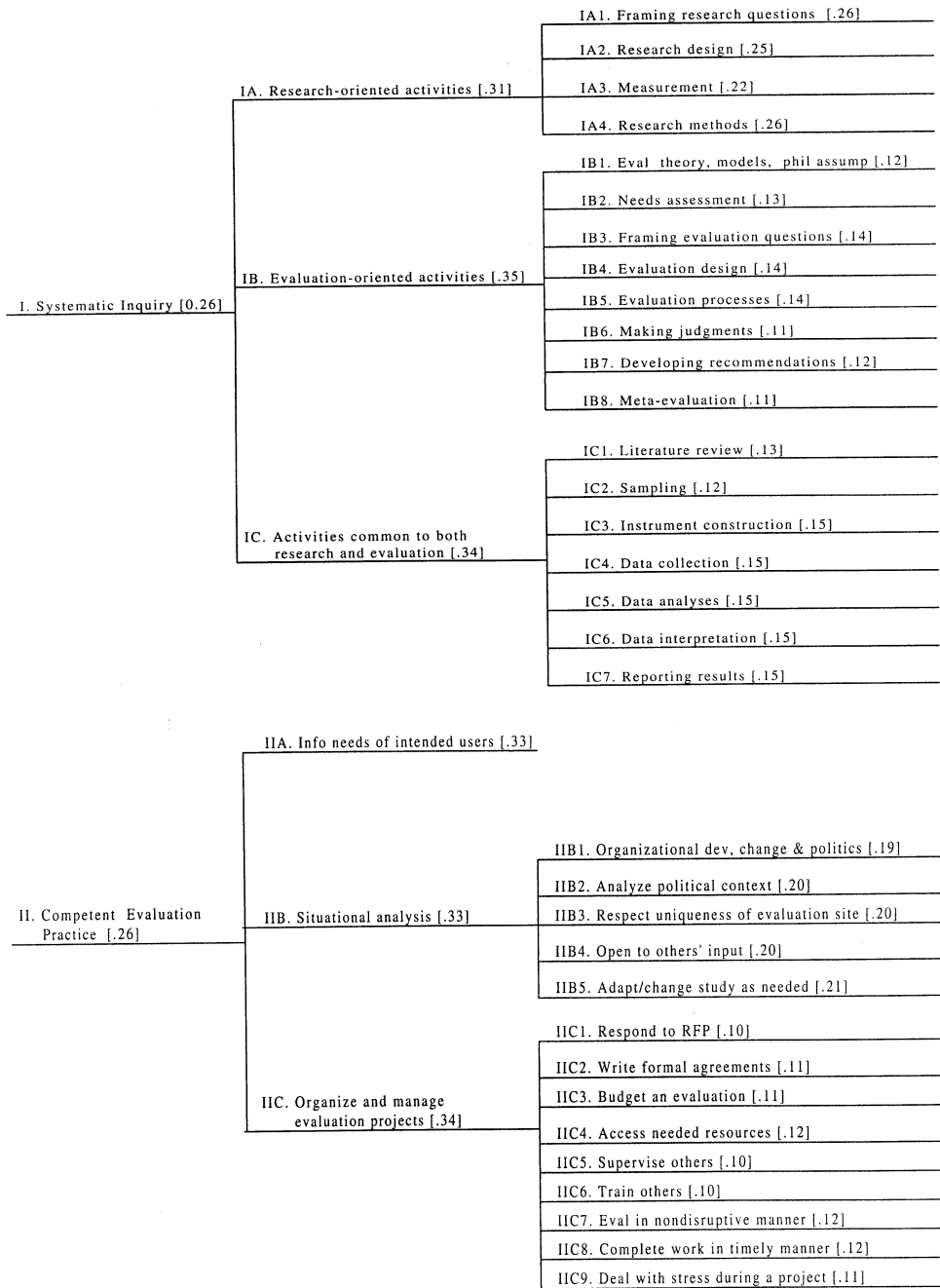


Figure 4. Weighted Importance of the Evaluator Competency Ratings.

needs to know more than research.” Another participant stated: “I would never frame a research question, ever. I only frame evaluation questions.” Participants generally accepted the idea that evaluators need to both know the difference between research and evaluation and be prepared to

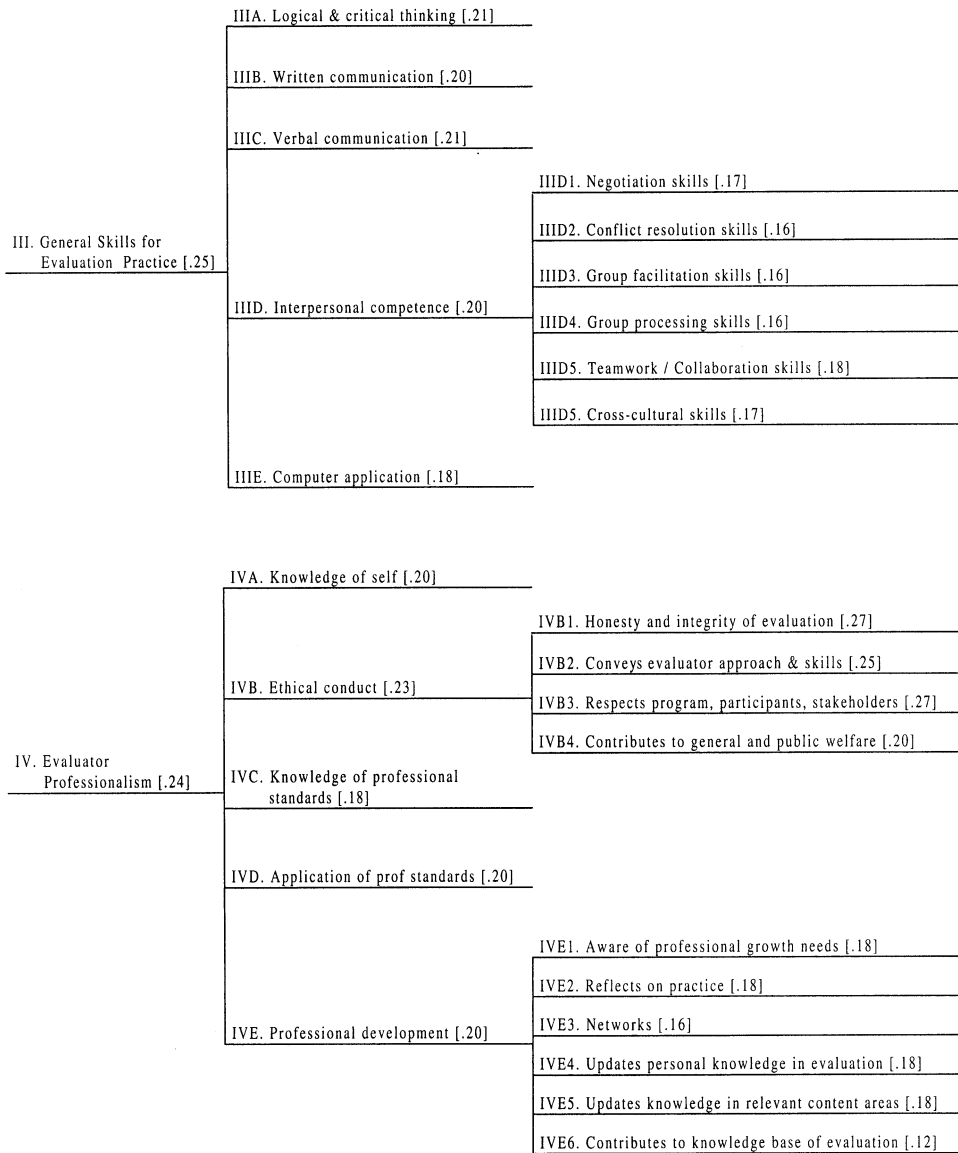


Figure 4. Continued

explain it to others. One participant expressed this view by saying: “I think it is important to explain that difference between evaluation and research to customers, because people hear evaluation [and] sometimes, when they are not familiar with it, think research.”

The second theme concerned the competencies that were of the highest perceived importance. Participants assigned the greatest importance to knowledge, skills, and attitudes for which evaluators had to take personal responsibility, that is, things that could not be hired out without affecting the integrity of the evaluation process. One participant, in discussing computer skills, said: “I rated [that] lower again because you could be the genius that thought

up [the evaluation plan] and designed it, but you do not have to be the person plunking away at the computer, or writing the stuff, or verbalizing it. So I rated everything except the thinker behind the scenes as lower." As another person noted, "I thought [that competency] was more a *nice to know* and less of a *need to know*." In discussing an evaluator's ability to organize and manage evaluation projects, one participant stated: "Well, if you can't do that, you can't do the evaluation."

Participants consistently focused on the impact of an evaluator's unique context by referring to the "real working world." Being grounded in the day-to-day reality of program evaluation influenced their perceptions about what competencies an evaluator actually needs to master, as opposed to those that might be helpful, but not essential. "Someone ought to do it, but in the world that I practice, I don't need that skill." Some participants clearly thought in terms of "What can be hired out?" or "At what point does an evaluator do a task herself or himself versus seeking assistance?" There was frequent dialogue that focused on the tug-of-war between the realities of the work world that make short cuts necessary and the belief that it would be good for evaluators to possess a wide range of skills, even those that could be contracted out to others. One person said, for example: "I encourage staff people I work with to do lit reviews or to include whatever they are familiar with to support their rationale for program choices . . . but it doesn't ever happen. It's not high on their priorities, so it's not high on mine."

CONCLUSIONS

What is reported in this article are the initial findings of an exploratory study concerning the face validity of a set of potential core competencies for the field of program evaluation. Humble practitioners that we are, we are well aware of the limitations of our work. Certain of the *Essential Evaluator Competencies* overlapped, making separate ratings difficult at best. Others contained terms that are difficult to define. Our sample was predominantly women, and some had limited evaluation experience. In addition, the relatively small sample size made subgroup analyses inappropriate. Our approach to establishing face validity of the competencies, MACR, had its own limitations, including the mandatory use of a 100-point rating as an anchor, the forced inflexibility of the competencies instrument we used, and the potential for groupthink and other social desirability effects. Furthermore, there was routinely too little time for complete discussion of each item during the sessions; people wanted to provide more detailed rationales for their ratings and to react to others' responses as well. The question of whether participants should respond from their own professional context as opposed to "evaluators in general" was routinely asked, and our response—speak from your unique experience—clearly affected the results. Like much in life, this was surely an imperfect process.

However, recalling the well-known difference between people who see a glass half empty compared to those who see it half full, we choose to see this glass half full. Despite its limitations, the face validation process also had its strengths: it was purposefully systematic; it generated both quantitative and qualitative data; and it used practicing evaluators with diverse characteristics who, therefore, constituted an inclusive "judging panel." In addition, participants reported they enjoyed the process, even though it was time consuming. What has become clear to us is that, despite their limitations, the data we collected suggest potentially useful conclusions that we present here to continue the field's ongoing conversation about

evaluator competencies. Specifically, we have reached the following conclusions after three years' work on this project.

There may be more agreement on the competencies needed by evaluators than one may have expected given the debate in the field to date. The 78% agreement on the competencies in this framework points to the fact that a diverse group of evaluators with a range of contexts and experiences, following a systematic rating and discussion process, found a great deal to agree on—certainly more than we would have predicted when we began. Further, the quantitative and qualitative data were aligned; the qualitative comments gave meaning to the quantitative ratings, suggesting the value of the mixed methods approach taken. It is important to note, however, the obvious need for further clarification of terms and concepts, even such basic ones as research and evaluation.

The areas where consensus did not emerge reflect the role- and context-specific nature of evaluation practice. The “real working world” in which participants grounded their evaluation responses surely influenced their ideas; what was required for a given role in one context (e.g., an independent consultant in the human services) made little sense for a different role in another context (e.g., an internal evaluator in a large government bureaucracy). The 15 identified areas of disagreement in Table 1 point to varying roles and contexts. For some people, knowledge of conflict resolution and cross-cultural skills was essential to effective practice; to others, these were not things to worry about. Some evaluators agreed with Michael Scriven that to evaluate is to judge; others did not. In developing a set of essential competencies for the field, the areas where no consensus existed may be the very areas where we might distinguish types of evaluation practice, acknowledge our differences, or create professional development opportunities.

The competencies that received the lowest ratings point to areas of continuing tension in the field, areas that deserve attention in professional development and training programs. The single item that received the lowest mean (60.84)—IVE6, “Contributes to the knowledge base of evaluation”—engendered passionate conversation. Even though evaluations do this (i.e., contribute to the knowledge base of evaluation) almost by definition, few participants believed this was part of their professional role. (Compare the role of a doctor or lawyer who uses the results of others' inquiry). Given the increasing ease with which evaluators can now share their work thanks to advances in technology and evolving methods for cross-study integration (e.g., meta-analysis, cluster evaluation), this is an area that may be worth noting. The other items with means below 80 were IB6, “Making judgments,” and IVB4, “Responsible for contributing to the general and public welfare.” The role of judgment in evaluation has been a source of debate for many years and clearly continues to be so. Disagreement over the evaluator's role with regard to the “general and public welfare,” however, may be more interesting because, always to participants' surprise, the language is taken directly from AEA's *Guiding Principles*, and yet few were willing to accept responsibility for the public welfare—willing to work on it, perhaps, but not to take responsibility for it.

This set of proposed evaluator competencies can be a helpful tool for professional reflection and discussion. Whether or not the field ever reaches full or even partial consensus about its competencies, the discussions held as we collected the data presented

here were often powerful. When people disagreed, they provided reasons and examples from their own practice and then listened intently to others with different ideas. Students and professionals alike became aware of areas they needed to work on or of types of evaluation practice that did not interest them.

Evaluation researchers need to conduct additional studies on potential competencies for the field. Such a statement is surely commonplace in the concluding paragraphs of virtually every research study. However, the initial results of the face validity of the *Essential Evaluator Competencies* suggest that what has traditionally been viewed as an impossibility may, in fact, be feasible. Additional studies should broaden to a national sample with substantial numbers representing diverse subgroups of evaluators and should address the conceptual problems related to overlapping competencies and definitions. Although MACR was a useful process for an exploratory study, researchers may want to consider other methodological options (e.g., Delphi techniques, surveys, or constructive controversy) as this work develops.

Worthen (1999) writes, "Evaluation competencies—skills and knowledge that enable an individual to conduct a quality evaluation study—represent the *sine qua non* in performance as an evaluator" (p. 546). As more career opportunities in the field of program evaluation create greater demand for preparatory and professional development programs, we believe it is time for those in the field of program evaluation to both identify and reach consensus on a set of core evaluator competencies. Doing so would provide a useful guide for establishing competency-based evaluation curricula both in higher education and professional organizational training programs. Such a guide also could serve as a useful tool for practicing program evaluators by enabling them to assess present levels of personal professional growth and to target skills and abilities for future development.

This project had extremely practical origins in an evaluation studies colloquium at the University of Minnesota where a group of evaluators new to the field sought to frame their own professional development. What began as a personal discussion evolved over time into the seeming search for our field's Holy Grail. Many discouraged us from the task—they said it could not be done—but virtually everyone who participated in our ongoing discussion did so vigorously. From the beginning, evaluation professionals and students engaged in lively conversation on these issues, repeatedly reminding us of the value of the discussion process itself. It is in this spirit that we present these initial results and invite others to join us in continuing the dialogue.

REFERENCES

- Altschuld, J. W. (1995). Developing an evaluation program: Challenges in the teaching of evaluation. *Evaluation and Program Planning, 18*, 259–265.
- Altschuld, J. W. (1999). The certification of evaluators: Highlights from a report submitted to the Board of Directors of the American Evaluation Association. *American Journal of Evaluation, 20*, 481–493.
- Altschuld, J. W., & Bickman, L. (1998). *Be it resolved that evaluators should be certified (credentialed): The pro side of the issue*. Paper presented at the annual meeting of the American Evaluation Association, Chicago, IL.
- American Evaluation Association, Task Force on Guiding Principles for Evaluators. (1995). Guiding

- principles for evaluators. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding Principles for Evaluators*. New Directions for Program Evaluation, 66 (pp. 19–26). San Francisco: Jossey-Bass.
- Anderson, S. B., & Ball, S. (1978). *The profession and practice of program evaluators*. San Francisco: Jossey-Bass.
- Covert, R. W. (1992). *Successful competencies in preparing professional evaluators*. Paper presented at the annual meeting of the American Evaluation Association, Seattle, WA.
- Ingle, M. D., & Klaus, R. (1980). Competency-based program evaluation: A contingency approach. *Evaluation and Program Planning*, 3, 277–287.
- Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Sage.
- Jones, S., & Worthen, B. R. (1999). AEA members' opinions concerning evaluator certification. *American Journal of Evaluation*, 20, 495–506.
- Joyce, B., & Weil, M. (1996). *Models of teaching* (5th ed.). Boston, MA: Allyn and Bacon.
- King, J. A., Minnema, J., Ghere, G., & Stevahn, L. (1998). *Evaluator competencies*. Paper presented at the annual meeting of the American Evaluation Association, Chicago, IL.
- King, J. A., Minnema, J. E., Ghere, G., & Stevahn, L. (1999). *Essential evaluator competencies*. Minneapolis, MN: University of Minnesota, Program Evaluation Studies.
- Mertens, D. M. (1994). Training evaluators: Unique skills and knowledge. In J. W. Altschuld & M. Engle (Eds.), *The preparation of professional evaluators: Issues, perspectives, and programs*. New Directions for Program Evaluation, 62 (pp. 17–27). San Francisco: Jossey-Bass.
- Merwin, J. C., & Weiner, P. H. (1985). Evaluation: A profession? *Educational Evaluation and Policy Analysis*, 7, 253–259.
- Patton, M. Q. (1990). The challenge of being a profession. *Evaluation Practice*, 11, 45–51.
- Sanders, J. R. (1979). The technology and art of evaluation. A review of seven evaluation primers. *Evaluation News*, 12, 2–7.
- Sanders, J. R. (1986). The teaching of evaluation in education. In B. G. Davis (Ed.), *Teaching evaluation across the disciplines*. New Directions for Program Evaluation, 29 (pp. 15–27). San Francisco: Jossey-Bass.
- Smith, M. F. (1998). *Should AEA begin a process for restricting membership into the profession of evaluation?* Paper presented at the meeting of the annual American Evaluation Association, Chicago, IL.
- Smith, M. F. (1999). Should AEA begin a process for restricting membership in the profession of evaluation? *American Journal of Evaluation*, 20, 521–531.
- Vanderwood, M. L., & Erickson, R. (1994). Consensus building. *Special Services in the Schools*, 9, 99–113.
- Worthen, B. R. (1975). Some observations about the institutionalization of evaluation. *Evaluation Practice*, 16, 29–36.
- Worthen, B. R. (1999). Critical challenges confronting certification of evaluators. *American Journal of Evaluation*, 20, 533–555.
- Worthen, B. R., & Sanders, J. R. (1991). The changing face of educational evaluation. *Theory into Practice*, 30, 3–12.