A Cultural Reading of
The Program Evaluation Standards
(2nd edition)

Final Report of the
Diversity Committee
American Evaluation Association

September 2004

Preface

This document contains the text conversation among a group of professional evaluators who share expertise in and concern for issues of cultural diversity and cultural context in evaluation. Some entries reflect individual viewpoints, and some represent the product of group discussions. For clarity, the symbol ◙ has been inserted to indicate a change of voice within a sub-section. The text has been edited to remove identification of individual readers and conversation among readers extraneous to the focus of the cultural reading. For clarity, cryptic comments such as, "See above" were spelled out, incorporating some purposeful redundancy. In no case was the meaning of a comment altered; differences of opinion among readers remain visible in the text. The following sections describe the background, purpose, and procedures of the reading, as well as membership in the Task Force.

Purpose. The purpose of the cultural reading is to review *The Program Evaluation Standards* (2nd edition) with respect to coverage of cultural diversity, treatment of cultural concerns, and attention to cultural competence in order to inform Diversity Committee members and other relevant stakeholders within the American Evaluation Association (AEA); to identify specific passages that are in need of revision; and to gather ideas on relevant resources and suggested changes for input to The Joint Committee on Standards for Educational Evaluation. **Culture is broadly defined, inclusive of race, ethnicity, gender, age, sexual orientation, social class, disability, language, and educational level or disciplinary background.** It includes both individual characteristics and those of a group or collective (e.g., community or organizational culture).

Background. At a face-to-face meeting of the Diversity Committee convened by Chair Edith Thomas at the 2002 Annual Meeting of the American Evaluation Association (AEA), discussion turned to agenda items for 2003. It was noted that *The Program Evaluation Standards* (2nd edition) would be up for consideration by the Joint Committee in fall 2003, and Karen Kirkhart expressed the strong opinion that these standards needed to be revised to update and improve their attention to dimensions of cultural diversity. This stimulated interest among members present. Kirkhart proposed that the Diversity Committee do a group reading of the *Standards* and record their critical reflections, exchange ideas, etc. so that the Committee would be in a more informed position to offer comment, reflect interpretations from different perspectives, and impact the process of *Standards* revision through established channels, via the AEA representative to the Joint Committee, Dianna Newman. The cultural reading became a 2003 action item for the Diversity Committee, chaired by Satish Verma. Karen Kirkhart volunteered to serve as lead person. The Board approved this action item at its February 2003 as an element of the Diversity Committee's 2003 Action Plan

Participants. Initial readers were members of the Diversity Committee in 2002 (when the idea was initiated) and 2003 (when Draft 1 was completed). Charles Thomas joined the reading group in Spring 2003 to provide continuity with the Building Diversity Initiative's Task Force on Guiding Principles for Evaluators Working Across Cultures, which he chaired. In alphabetical order, the initial readers were:

Claude F. Bennett, U.S. Department of Agriculture
Denice A. Cassaro, Cornell University
Melvin E. Hall, Northern Arizona University
Stafford Hood, Arizona State University

Lester Horvath, Evaluation Consultant, Connecticut
Elmima C. Johnson, National Science Foundation
Karen E. Kirkhart, Syracuse University
Donna M. Mertens, Gallaudet University
Sharon Rallis, University of Connecticut
Edith P. Thomas, U.S. Department of Agriculture
Charles L. Thomas, George Mason University
Satish Verma, Louisiana State University
Elizabeth Whitmore, Carleton University

The work of the 2003 Task Force was supported by:

Crystal Collette, Syracuse University
David Schlesselman, Syracuse University

In 2004, the Diversity Committee, chaired by Melvin Hall, appointed a Task Force to synthesize the text generated by the initial readers and to prepare comments in formats suitable for dissemination to the Joint Committee and other relevant audiences. The members of the 2004 Task Force were (in alphabetical order):

Denice A. Cassaro, Cornell University
Cindy A. Crusto, Yale University
Melvin E. Hall, Northern Arizona University
Elmima C. Johnson, National Science Foundation
Karen E. Kirkhart, Syracuse University
Joan LaFrance, Mekinak Consulting
Donna M. Mertens, Gallaudet University
Craig W. Russon, Kellogg Foundation.

Procedure. The first round of the Cultural Reading began in January 2003 and concluded in August 2003, yielding roughly 100 single-spaced pages of text material, summarizing and critiquing the *Standards*. Members read the standards together, posting their comments to the entire group via email, then conversing and exchanging ideas around the postings. Readers created their own "archives" by saving the emails to a folder. Each standard was read in order of presentation in the book. Not everyone logged in comments on every standard, and some readers were more active than others in the process. Karen's Spring 2003 Graduate Assistant, David Schlesselman pulled together the comments at the end of each of the four sections—Utility, Feasibility, Accuracy and Propriety—as well as an integrated document at the end. This became Draft 1. In Fall 2003, Crystal Collette, then a Graduate Assistant at Syracuse University, searched the literature and compiled a bibliography of culturally relevant references by standard. This was provided to the Joint Committee as a freestanding document in 2004 and incorporated in the second draft of the Cultural Reading.

In 2004, moving the cultural reading from a private to a public document remained an action item of the Diversity Committee. The second round of the Cultural Reading began in January 2004 and concluded in August 2004, yielding a set of prioritized action items and summary matrixes organized by individual standard as well as by category (Utility, Feasibility, Propriety, Accuracy), a narrative summary, and an edited version of the full text, identified here as Draft 3. The content and format of these products are as follows:

- ***Executive Summary***: Narrative summary, including overview of recommended corrective actions.
- ***Priority Recommendations***: Important corrective actions summarized for all standards, matrix format.
- ***Standards Overview by Categories***: Summarizes the relevance to cultural competence, current strengths, concerns or limitations, and recommendations for standards by major category—Utility, Feasibility, Propriety and Accuracy, matrix format.
- ***Summary of Utility Standards***: Summarizes the relevance to cultural competence, current strengths, concerns or limitations, and recommendations for each of the Utility standards, matrix format.
- ***Summary of Feasibility Standards***: Summarizes the relevance to cultural competence, current strengths, concerns or limitations, and recommendations for each of the Feasibility standards, matrix format.
- ***Summary of Propriety Standards***: Summarizes the relevance to cultural competence, current strengths, concerns or limitations, and recommendations for each of the Propriety standards, matrix format.
- ***Summary of Accuracy Standards***: Summarizes the relevance to cultural competence, current strengths, concerns or limitations, and recommendations for each of the Accuracy standards, matrix format.
- ***A Cultural Reading of the Program Evaluation Standards (2nd ed.)***: Narrative discussion among readers of each of the thirty standards, inclusive of Overview, Guidelines and Common Errors, Case Illustrations and Analyses, and suggestions for Supporting Documentation.

Next Steps. This document was approved by the Diversity Committee on September 28, 2004 and by the AEA Board of Directors, November 3, 2004. These documents are available to support the work of AEA members, other AEA Committees and the Task Force of the Joint Committee on Standards for Educational Evaluation charged with revising the second edition of *The Program Evaluation Standards* (Donald Yarbrough, Chair). Elmima Johnson, a participant in both phases of the cultural reading, assumes the position of AEA representative to the Joint Committee in January 2005. The Diversity Committee will seek continued opportunity for input as the Joint Committee Task Force moves through the revision process. We welcome dialogue and discussion. Comments may be addressed to the Chair of the Diversity Committee, Dr. Melvin E. Hall, the Chair-Elect of the Diversity Committee, Dr. Joan LaFrance, or the Chair of the Cultural Reading Task Force, Dr. Karen E. Kirkhart.

Return to Table of Contents

Table of Contents

Utility Standards

The Utility Standards are intended to ensure that an evaluation will serve the information needs of intended users.

U1 Stakeholder Identification.

Persons involved in or affected by the evaluation should be identified, so that their needs can be addressed.

Standard.

This is an important standard, one that is highly relevant to culturally competent evaluation. It should be retained. While the wording is very general (as is true of all of the Standards) I see no need for a change.

Overview.

I like it that the first sentence makes reference to diverse stakeholders, even though it does not explicitly refer to cultural diversity. The remainder of the first paragraph focuses on the variety of roles that may come into play, which is fine. I personally feel the list of roles is top-heavy with management/administration, and I would like to see a bit more elaboration on the consumer side to add balance. For example, it's not just "consumers who purchase goods and services" but consumers who are mandated into programs, those who are in need of goods and services, families of consumers, and neighborhoods and communities that may be impacted by program interventions.

◙I support this comment and also feel that the term consumer needs either elaboration or to be joined by additional words which describe the other non-consumer relationships various stakeholders may have with a program or product. It is hard to think of a participant in a mandated program as a consumer who has purchased goods and services.

Additionally, I believe that the concern you outlined about priority of stakeholders is created by the use of "furthermore" in the last sentence of that first paragraph. This phrasing seems to relegate the "typical stakeholders" listed as belonging to a less central group than the earlier listing. I would propose rephrasing the opening of that last sentence to avoid this perhaps unintended ranking of stakeholders.

◙It's really the second paragraph of the Overview that I think could be strengthened. The second sentence doesn't sit quite right with me. "In many evaluations, special efforts may be necessary to promote the appropriate inclusion of less powerful groups or individuals as stakeholders, such as racial, cultural, or language minority groups." First, it feels condescending to me, as if these groups have less power by definition rather than as a result of majority oppression. Second, it seems to imply that this is a special case, not true of all evaluations. My view would be that all evaluation is culturally contextualized; therefore, one would always include consideration of cultural dimensions in the identification of stakeholders (though it wouldn't necessarily always be race or language that was the defining characteristics of relevance). Third, the dimensions of cultural diversity should be expanded so that the reader is prompted to consider variables such as economic status, ethnicity, education, sexual orientation, age, disability, religion, gender, health status, immigration status and to understand that cultural considerations are not narrowly defined.

◙Again, I support this perspective.

Guidelines.

(A) raises a good point, but could be further strengthened by mentioning community leadership or something that would make clear that it's not just authority figures within organizations that should be identified.
◙Would altering the opening phrase to: Identify persons in formal and informal leadership roles" help with this issue?

◙ (B) raises the important issue of how to identify representatives of stakeholder groups. This is especially important when the evaluator may be an outsider to the group from which a representative is sought. Do we have any good references here? Avoiding tokenism in evaluation?
(D) operationally shouldn't this one read "Reach an **initial** understanding with the client concerning the… otherwise E cannot really happen.

(G) Good that nondiscrimination is explicitly stated. As above, I would expand the list of illustrative dimensions.
Common Errors. I like what is here, though I would expand the descriptors under (F).
It seems like tokenism might be an error worth adding to the list. I think that evaluators sometimes settle too quickly for "representation" without considering who the representative is or what role that person is to play.
I would like to see some consideration given to the error of "failing to anticipate competing or adversarial views of program goals and objectives held by stakeholder groups." It becomes impossible to honor these different views in the evaluation when the evaluator is totally unprepared to acknowledge that many program goals are contested along stakeholder-group lines.

Illustrative Case 1 – Description. As in most of the case studies, there is no mention of cultural diversity (except age in this case, as it is an early childhood program), leaving the impression that such characteristics are not relevant to consider. Here, for example, a sentence about the economic status of the community or of language diversity present would appear to me to be relevant to the story. (I think we should be careful to infuse diversity in relevant ways, not as some politically correct litany of categories that could easily be dismissed.) It is not clear to me how closely the Illustrative Cases follow the facts of an actual event, but if it were reality-based, then the author would be able to judge the dimensions to include. I also felt that the mention of "early childhood interest groups" (whatever those refer to) at the very end seemed dismissive of their input as an implied partisan perspective. The focus of the case is stakeholders at the upper levels of administration and legislative bodies.
Illustrative Case 1 – Analysis. The analysis, while not explicitly addressing cultural dimensions, does a nice job (paragraph one) challenging the authority-driven definition of stakeholders and emphasizing a more balanced perspective.
Illustrative Case 2 – Description. The case has so many problems that the Stakeholder Identification piece sort of gets lost. It would seem to illustrate a Violation of Information Scope and Selection (U3) better than U1. But here again, if it were edited, relevant cultural factors should be introduced. For example, "factors that would influence placement rates" might include racism or sexism in the workplace.
Illustrative Case 2 – Analysis.
◙ I don't disagree with the author's conclusion that the results were invalid, but I still don't see Stakeholder Identification jumping out as the source of the problem. They knew

who the stakeholders were; they just didn't act on it in developing the design. If the Joint Committee is open to replacing some of the case illustrations in the revision, I'd flag this one to be replaced. Neither case illustration tackles the tough issue of deciding who speaks for stakeholder groups outside of hierarchical organizations or systems.

◙ I am troubled by the use of this case if there is a failure to clearly pull out several key factors. First, the absence of any attention to the actual participants who went through the programs and were placed is a concern. If program effectiveness and outcomes were the goal, failing to identify those who participated as potential stakeholders is a major limitation of the study. Also the rather loose presentation of this case fails to clearly identify the charge to the evaluation. Was it appropriate to operationally define program success as the number of graduates from college programs who found jobs, how soon after graduation they were employed, and how long they kept their positions?

And as a final point, the problem with using cases in this format is that it suggests the only problem with the case is the one that is the focus of that standard. At a minimum the final sentence might allude to these issues as limitations.

◙ I do agree that the predominate omission appears to be in the conceptualization of participant groups as stakeholders. In the guidelines, cases and case analyses, either they are ignored, presented as less important or given "token" attention. And little specific attention is given to potential problems in stakeholder identification and participation, when they are "not of the majority."

Supporting Documentation.

◙ I would nominate Mathie & Greene (1997) – though we may find it fits better under another standard. It's one of my favorite articles for taking a tough look at diversity and inclusion.

Gilliam, A., Davis, D., Barrington, T., Lacson, R., Uhl, G., & Phoenix, U. (2002). The value of engaging stakeholders in planning and implementing evaluations. *AIDS Education and Prevention, 14*, 5-17.

Mathie, A., & Greene, J. (1997). Stakeholder participation in evaluation: How important is diversity? *Evaluation and Program Planning, 20*, 279-285.

Mercier, C. (1997). Participation in stakeholder-based evaluation: A case study. *Evaluation and Program Planning, 20*, 467-475.

Thayer, C., & Fine, A. (2001). Evaluation and outcome measurement in the non-profit sector: Stakeholder participation. *Evaluation and Program Planning, 24,* 103-108.

Return to Table of Contents

U2 Evaluator Credibility.

The persons conducting the evaluation should be both trustworthy and competent to perform the evaluation, so that the evaluation findings achieve maximum credibility and acceptance.

Standard. Evaluators and stakeholders are the principal parties in evaluation, and both standards (UI and U2) are important and relevant to the conduct of evaluations that are culturally relevant and meaningful. The wording of this standard is appropriate and stresses that competence and trustworthiness of evaluators are important for findings to be utilized by stakeholders/primary intended users.

◙ This is an extremely important standard, one that is central to our concerns. General understandings of what it means to be "competent to perform evaluation" must be broadened to include cultural competence. Even more significant to my mind is that this Standard potentially unpacks prejudice surrounding "credibility." As I will argue below, in many academic environments, "lack of credibility" is a thinly veiled euphemism for "not like me."

Overview. Criteria for evaluator credibility are rightfully indicated in the opening paragraph. In the opening sentence of this paragraph, I would add "cultural awareness and sensitivity" after public relations skills. In the same sentence, "other characteristics" is too general to mean anything, and the qualification that these other characteristics are "considered necessary by clients and other users of evaluation findings and reports" could be difficult to evaluate and implement. The second sentence of this paragraph makes the useful point about the need for a team of persons (since no one individual can possess all credibility characteristics) to do an evaluation. Diversity and inclusiveness of the evaluation team should be desired.

◙In the first paragraph, I like the recognition of personal limitations and the importance of constructing a team that collectively possesses the needed qualities. As others have already noted, competence must be broadened to include more than technical competence. Cultural competence should be explicitly noted here.

◙Although I am not sure how to fix the problem, the first paragraph of the overview seems to be a parallel to "face validity" in a measurement context. Shouldn't this be a combination of appearance and performance with more weighting on performance?

◙Again the importance of participant views is ignored.

◙In the second paragraph, while I agree that credibility must be addressed from the beginning, I think it needs to be clear that this is not a simple matter of "first impressions." Credibility may be gained or lost at the outset *or at any time during the evaluation process*. I agree with the strong statement of its importance, but I think the Standard should be clear that this is a pervasive concern throughout the evaluation process, not just a "front-end" issue to be resolved so that the evaluation can move forward.

◙I would add stakeholders to the first sentence of the second paragraph: "Evaluators should establish their credibility with the client, stakeholders, and other users at the …" The intent is to link with the broader identification of stakeholders in a revised Standard One.

◙2nd paragraph- "If they do go ahead when they are considered unqualified or biased...." This sentence is ludicrous. Why would a client allow them to proceed under these circumstances?

◙The second and third paragraphs could be modified to include the idea of cultural competence of evaluators in "reading" the evaluation context and audiences to engender trust upfront (paragraph 2) and maintain communication and approachability with clients and stakeholders.

◙Last sentence - "test of credibility is the ability to "defend"... Enron executives were able to defend their actions until a whistleblower came forward. The test is the actual utility, integrity, etc., not the ability to convince people that it has these qualities.

◙This third paragraph is a telling reflection of its author’s position. The list of what the “fundamental test of their credibility will rest in” reflects an academic, social science perspective. These factors may not be the key credibility concerns of all audiences, nor do they include fundamental culturally-relevant issues such as historical legacy and respect.

Guidelines. (A) makes the useful point about staying abreast of social and political forces associated with the evaluation. To the list of forces mentioned could be added other dimensions of diversity, i.e., education, sexual orientation, age, disability, religion, health status, immigration status. ◙ (A) I like it that cultural dimensions are addressed first, but “staying abreast” is not a very strong mandate. I’d like to see a stronger statement here, as well as additional cultural dimensions added to the list—ethnicity, disability, sexual orientation, age. ◙Would it be useful to differentiate between “social and political forces” which to me connote less relevant features of the evaluation, and issues arising from factors that impact how program outcomes are valued by stakeholder-groups? It is the knowledge of how various stakeholders value key aspects of the program, which legitimates the claims of those who hold different cultural perspectives.

◙ (B) – (E) appear relevant procedures to implementing this standard. ◙ (B) “and the cultural context of the evaluation” should be added, citing A2.

◙ C - Have the evaluation plan and work reviewed for “cultural sensitivity” by members of the participant group, other than the team members. ◙ © I like the inclusion of Meta-evaluation, citing A12! I agree that when used to gain a cultural perspective or to provide a cultural critique, meta-evaluation can enhance the multicultural validity of evaluation.

◙ D - ...technically sound “and appropriate for the cultural context of the study” ◙ (D) “and culturally valid” should be added.

(E) is an appropriate guideline, but the examples are all pretty mainstream. I’d like to see greater emphasis on tailoring communicative devices to fit the audience, with at least one example that’s designed to get folks thinking, “Now that *would* require a different approach!”

(F) I’d like to see “qualifications” include personal characteristics and lived experience as they apply to the evaluation in question. ◙ (F) Remember to include cultural competence in evaluator’s qualifications.

◙*Common Errors*. (A) should include cultural and experiential areas as dimensions of relevance to credibility. ◙ (A) Rephrase “Failing to establish the evaluator’s credibility in content, methodology, and cultural competence”. ◙ A. Add cultural competency to the list.

◙ (B) is curious to me. I agree that bias is important to address, but as written I’m not quite getting how this is a matter of *evaluator* credibility. I guess to do so would reflect poor judgment on the part of the evaluator (though that could apply to violation of almost any Standard). Also, it’s not clear where the client falls in this statement. If the client is included as a stakeholder, then I would agree with the statement, but if a contrast is being suggested that positions stakeholders as “Other,” I would find this too narrow. I would suggest that matters of perspective and potential bias should *always* be thoughtfully examined. Also, I am not clear where advocacy models of evaluation fall here. (But Stufflebeam’s vocal opposition to such models makes me watchful.) I would argue that the Standards should be written in such a way that they support all models of evaluation, not rule some out by definition. ◙ B. What does this mean?

◙ © I think is OK as long as cultural competence would be included among relevant skills and experience. ◙ C. Add cultural competency to the list.

◙ (D) It's interesting to notice that here the examples have explicitly moved beyond education, because that has been a source of debate concerning the Standards over the years—Are they education-specific? This relates to the appropriate composition of the Joint Committee itself. I would like to see some non-mainstream examples added—again in the sprit of getting readers to reflect on more diverse settings; e.g., American Indian reservations, migrant camps, homeless shelters, community centers, prisons. ◙ D. Again not just setting but cultural context as well. Also for NA groups, geography itself is an important variable.

◙ (E) I'm not quite sure what it means to "devote…their reputations" but the general principle of making a commitment to the study is good. . ◙ E. How do you "devote your reputation?"

◙ (F) I'm not really sure what this refers to, but it seems that it *could* be used to argue against investing necessary time to gain entry to a community or setting, and this would not be desirable. Establishing credibility across dimensions of difference is often an extensive, slow, labor-intensive process, and I would hate to see such investment cast an Error. (G) Seems OK. (H) again raises the issue of how the client is positioned (see my comment on Guideline B above), but certainly issues of value differences between client and evaluator are important to recognize.

◙ I. Does this need to be said? Also you don't have to be a student to be "inexperienced." ◙ Does (I) happen in real life? I would suggest taking out the word student and just say inexperienced assistant. There is some "bait and switch" activity where a name evaluator is hired but really does not do the work or closely supervise those on the job. ◙ (I) I agree with previous comments that it seems unnecessary and inappropriate to single out "student." The Error could apply to anyone in an assistant capacity that lack experience.

Illustrative Case 1 – Description. This case does allude to cultural diversity but only in the non-specific terms, "minority" [students], and "minority group members" [law faculty]. First interesting point is what dimension(s) of diversity came to mind; I thought of race (which itself may be racist but so goes this tangle). We are explicitly told that the students are second class citizens in this Academy, having scored poorly on traditional admissions measures, which is attributed to their lack of "basic communication and study skills" (perhaps a questionable assumption, but less relevant to this particular Standard). The way in which the text introduces the faculty "themselves minority group members" also sets them apart from the mainstream, and their concerns regarding cultural barriers—which they may be uniquely qualified to observe and understand—is "politely but firmly rejected," [what condescension!] because it seemed clear to the assembled majority that these law professors' role was illegitimate. Wow! Now we notice that they base their assertion of illegitimacy on the fact that this evaluation was not formally commissioned by a person in a position of higher authority (i.e., the professors did not ask permission to conduct this formative, internal evaluation) and that the respondents were volunteers (a typical procedure for protection of human subjects). A potentially more relevant methodological concern might illustrate Information Scope and Selection (U3) or

Defensible Information Sources (A4), but the broader point here is that this convened group of academic colleagues never had any intention of taking the concerns of these minority faculty seriously, and they were seeking acceptable ways to discredit them and their effort, within the culture of the Academy (challenge their authority; attack their science). A powerful case illustration to be sure, but not in the direction taken by the analysis.

◙ Both cases are relevant to the standard. Case 1 addresses the issue of cultural competence and sensitivity and could be retained since it brings out the important point of lack of evaluator credibility.

I continue to be concerned about the loose presentation of cases. For the first case, did the professors say they were doing an evaluation of the materials? It appears that they already had a position which they sought evidence to support. This point is critical because they are later indicted for not having evaluation expertise but cited for their knowledge of minority affairs. If I were one of the faculty members I would wonder why my expertise as a law professor was not sufficient grounds for credibility in evaluating law school curricular or remediation materials. I think this case needs considerably tightening to be useful and not a negative example with subtle judgments sending a message contrary to inclusiveness.

◙ Case 1 - Racist. I don't know where to begin...with the characterization of the minority law professors as bumbling, well meaning idiots or the elitism of the chairs who "politely rejected the recommendations" or the hidden agenda of the Dean who had to know it would be rejected..... A more realistic example would be to have external evaluation experts perform the evaluation and conclude there were no barriers because of their lack cultural competence.

Illustrative Case 1 – Analysis. The analysis begins with a helpful clarification that the deficiencies identified by the Professors were real, that they had not been previously recognized by the designers of the materials, and that should be corrected ("taken into consideration" describes weak action with little accountability for follow-up) in the design of future materials. Astonishingly, however, rather than taking up issues of institutional racism and the dynamics of achieving "legitimacy " in the Academy, the analysis proceeds to fault the Professors for "failing to address their credibility in the design and conduct of the evaluation." It's not clear exactly what the analyst had in mind as appropriate steps to insure credibility beyond seeking the approval of the Dean, peers, and stakeholders, but based upon my experience with Academe, that would not have addressed the underlying issues. Aspersion is cast on their expertise as "knowledge of minority affairs" which is explicitly set outside the parameters of "evaluation expertise" for which they should have gone to the very unit that had demonstrated cultural insensitivity in its materials, the Teaching/Learning Center! (ignoring the Conflict of Interest (P7) violation that this advice could create.) There is real potential in this case analysis to take up the question of when "credibility" is a code word for racism, sexism, heterosexist bias, but that is never even alluded to as a possibility. At best, this case could be used to illustrate Political Viability (F2) because the Professors clearly lacked political clout in this system, but to allow it to stand as an example of personal credibility of the Professors is to be complicit in a racist dialog.

Illustrative Case 2 – Description. The disrespect shown the minority faculty in the first case stands in stark contrast to the respect shown the medical students in the second case, although this case has its own problems. Here, technical competence is conflated with credibility. If the questionnaire were poorly worded, the response rate insufficient, the scope of the study too restrictive, these are all methodological points that could be better used to illustrate Accuracy standards such as A5 Valid Information. Standard U2 is written to address evalua*tor* credibility, not evalua*tion* credibility, so the concerns do not seem to be a particularly good fit.

◙ Case 2 appears to have been included to demonstrate flagrant violation of this standard in that students assigned to do the evaluation lacked any of the necessary qualifications and were destined to show their biases in the findings. I think it needs to be replaced with a better example.

◙ Case 2 - Please...What medical schools take complaints from students seriously!!!

Illustrative Case 2 – Analysis. Evaluators were members of "a medical fraternity at a large Midwestern medical school." No further information is given on the diversity within this population, which I understand to be male based on the "fraternity" designation. The analysis points to the fact that the students were "not sufficiently equipped" to conduct the needed evaluation. Methodological faults are referred to, citing A5; I agree with this analysis. Without discussion of what qualities made the external evaluation specialist credible, this second illustrative case doesn't contribute much too understanding standard U2.

I'm fairly certain that these cases were written and analyzed separately, so the author of this analysis is not responsible for "order effects" flowing from Case 1, but the contrast is striking. In the second case, a fatally flawed evaluation was taken as a "catalyst" for reviewing the evaluand, and students on the fraternity committee (the evaluators in the original study) were respectfully included in the redesign and expansion of their initial work. One wonders how issues of power, privilege, academic culture (medical school versus law school perhaps), and personal characteristics of these students (e.g., economic status, social class) shaped these two very different responses.

Supporting Documentation. All references are on internal evaluation and internal evaluators. I suspect this is because internal evaluations/evaluators are seen in a less favorable light than external evaluations/evaluators on the question of bias, although I doubt the veracity of this position. Regardless, other references should be added, especially those dealing with cultural competence. I believe the guiding principles for evaluators and commentaries written about them would be helpful. *New Directions for Program Evaluation*, Volume 66 may be a good addition.

◙ I would nominate Rodney's [Hopson] NDE volume on Language Matters here, engaging in a deconstruction of the term credibility in the same way that Anna [Madison] did for "at-risk."

Hopson, R. K. (Ed.) (2000). *How and Why Language Matters in Evaluation, New Directions for Evaluation*, No. 86, San Francisco: Jossey-Bass.

Madison, A. M. (1992). Primary inclusion of culturally diverse minority program participants in the evaluation process. In A. Madison (Ed.), *Minority issues in program evaluation, New Directions for Program Evaluation*, No. 53 (pp. 35-43). San Francisco: Jossey-Bass.

Madison, A. M. (2000). Language in defining social problems and in evaluating social programs. In R. K. Hopson (Ed.) *How and Why Language Matters in Evaluation, New Directions for Evaluation*, No. 86 (pp. 17-28). San Francisco: Jossey-Bass.

Quintanilla, G., & Packard, T. (2002). A participatory evaluation of an inner-city science enrichment program. *Evaluation and Program Planning, 25*, 15-22.

Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, C. (Eds.) (1995). *Guiding Principles for Evaluators, New Directions for Program Evaluation*, No. 66. San Francisco: Jossey-Bass.

Smith, L. T. (1999). *Decolonizing methodologies: Research on indigenous peoples*. New York: Zed Books, Ltd.

Wallerstein, N. (1999). Power between evaluator and community: Research relationships within New Mexico's Healthier Communities. *Social Science Medicine, 49*(1), 39-53.

Return to Table of Contents

U3 Information Scope and Selection.

Information collected should be broadly selected to address pertinent questions about the program and be responsive to the needs and interests of clients and other specified stakeholders.

Standard. The standard is an extremely important one, as it defines the parameters of inquiry, specifies voice and evidence. Any move to include cultural variables and perspective in an evaluation would call upon this Standard. The standard itself seems appropriately written.

Overview. The overview explicitly cites multiple stakeholders and the importance of opportunities for input, but then moves to illustrate additional variables that the evaluator should strive to include, whether or not they are nominated by stakeholders, in the spirit of including "all important variables." I concur with this perspective; overriding moral, legal, or ethical dimensions should necessarily be considered. I would nominate "equity issues" or similar variable ("fairness", "social justice") to the illustrative list. This perspective raises interesting issues of power and ownership of the evaluation design and who shapes it. In the second paragraph of the overview, the evaluator is charged with making the judgment of what is minor (to be discarded) and what is major (to be emphasized). While I think this is appropriate as a general statement of responsibility, I wonder what the Theory TIG would say if they were doing a theoretical reading of the Standards. It seems to me that some models of evaluation would reject vesting this much power in the evaluator. Good those stakeholder perspectives are repeatedly mentioned as part of the "weeding out" though.

Interesting that the caution (p. 38) on not letting testing (as an example of a mandated evaluation procedure) drive curriculum (as an example of practice) is now violated as a matter of National Policy. This is an issue of particular relevance to cultural dimensions of race, ethnicity, economic status, and language, though culture is not explicitly mentioned.

The process description (p. 38, paragraph 2) seems appropriate to me, though where it says, “This is done to ensure that the information to be collected addresses the important issues” I would add “and is culturally relevant.” ◙ 2nd paragraph, third sentence, “...strives to assess the program in terms of... (add cultural responsiveness)”

◙ In the last paragraph, the statement that evaluators “bring their own preferences” seems too soft. I would favor expanding that to emphasize that evaluator preferences are shaped by life experience, academic training, cultural identification, and area of practice. I agree with cross listing Values Identification (U4) here, and I would also add Meta-evaluation (A12) to support the point made previously about the opportunity for stakeholder review of the evaluation plan. ◙ 3rd paragraph ..Share the evaluation plan prior to data collection...” not just to address important issues but also to assess its cultural relevance to participants.

Guidelines. What strikes me about these Guidelines is their cut and dry, formulaic view of synthesizing and selecting evaluation questions. Such procedures may not result in the most culturally relevant questions being included. Specifically, I would add “A. Understand the cultural context of the evaluation (see A2 Context Analysis).” (B) presumes that interviewing is the appropriate way to gain understanding of the points of view of major stakeholders. This seems too narrow, given the range of strategies by which evaluators can become informed about diverse worldviews.

E) puts the power of ranking the importance of potential audiences in the hands of the client. Here again, this is congruent with some, but likely not all, models of evaluation. The criteria for such ranking should be spelled out and examined for potential bias, including culturally-based bias. ◙E. Client rank of importance of audiences may be biased or uninformed. The evaluator has a responsibility to broaden the client’s perspective as appropriate.

◙ Under (H), although this appears logical on first reading, I wonder how it would play out, say, against a culturally-relevant mid-ranked question that requires more work to answer (e.g., because of personal contact, trust building, etc.). In other words, I would assert that working across cultural boundaries to answer evaluation questions may require a more labor-intensive, time-intensive effort. If the level of evaluation effort is distributed only with consideration of importance rankings, as (H) suggests, it might create a systematic gap insofar as the culturally-relevant questions never get allocated the resources necessary to answer them well (unless they are top-ranked). ◙ H. There is not necessarily a 1-1 correlation between rank of items and effort required in each component of the evaluation.

Common Errors. (B) is on the right track. I’d add “and cultural perspectives” after “multiple stakeholder groups” to underscore cultural relevance here. While I respect the intent of (C)—updating information contacts—I think the strategy described (“periodic contacts”) is too limiting. I’d be more comfortable with a statement indicating that the best way to maintain an awareness of shifts in information requirements or other evaluation-relevant issues will vary with the stakeholder group, and procedures should be followed that are congruent with and respectful of the norms of each group.

Illustrative Case—Description. No mention of the cultural composition or location of the district, though it is apparently one in which children walk to school, based upon the comment on p. 40 concerning safe escort of young children. Also missing from the description is any mention of the superintendent’s avowed purpose for requesting the report against which to compare the choices made by the panel. Since no single evaluation can address all potentially relevant questions, it's difficult to judge the wisdom of the panel's actions without hearing the charge.

◙ The issues here are bias and credibility of the evaluators. Information scope and selection are secondary. Another unidentified issue is the purpose of the evaluation. Without that piece of information the effort was doomed to fail. Cultural factors, especially the age of the students and what that meant for the parents also was ignored. I could go on listing the inappropriateness of this example, but I think I've made the point.

Illustrative Case—Analysis. This analysis appears to offer a 'textbook perfect" answer that may not have fit the circumstances. I agree with the analyst's view that the time frame was inadequate to the level of analysis that was desired. I think we may want to consider a Standard that addresses time, beyond the reporting sense that we'll see in U6. I keep reflecting on how time is a validity threat insofar as there is often insufficient time to do the front-end relationship-building necessary to support multiculturally valid evaluation. Even though this example does not address culture, the analyst's recommendations are predicated on the assumption that the superintendent would extend the time frame. This may or may not be true. There is room here to add a second Illustrative Case that draws out cultural dimensions of Information Scope and Selection more clearly.

Supporting Documentation. I don't have any immediate specific suggestions here, though I notice that those used last time were very generic texts, with the exception of Stecher & Davis (1987). I would think that there would be some relevant content in the cultural competence literature that we could apply here, something touching on issues of power and privilege in building a relationship and developing a focus. Also maybe something from Patton, whose Developmental Evaluation presents a different view of how and when evaluation questions are set and prioritized—a counterexample to the formulaic, front-end approach presented in U3.

Cockerill, R., Myers, T., & Allman, D. (2000). Planning for community-based evaluation. *American Journal of Evaluation, 31*, 351-357.

Green, B., Mulvey, I., Fisher, H., & Woratschek, F. (1996). Integrating program and evaluation values: A family support approach to program evaluation. *Evaluation Practice, 17*, 261-272.

Return to Table of Contents

U4 Values Identification.

The perspectives, procedures, and rationale used to interpret the findings should be carefully described, so that the bases for value judgments are clear.

Standard. The standard itself is clearly written at a general level. The only possible change I might make is to add "standpoints" to the list of descriptors to make reference to culturally defined perspectives. This standard is extremely relevant to cultural diversity and evaluation. ◙Actually this standard needs to be rewritten. It leaves the impression that values are only important in the interpretation of findings, versus the entire process.

Overview. ◙ This standard reflects the core of cultural competency, because it is values, more than knowledge that determines cultural understanding and therefore competency. Even if the approach is agreed upon, the values assigned may differ. ◙While the opening paragraph is clear, it seems written at a pretty low level for professional evaluators. I guess I would prefer a bit more theory here, connecting values to the logic of evaluation. This is a general comment, however, not anything specific to our cultural reading. In the second paragraph of the overview, the reference to deciding who will make the value judgments and determining what procedures they will use could be expanded to point out issues of power surrounding values identification more explicitly. I would favor a stronger closing statement in paragraph three concerning the centrality of values identification to the entire evaluation process, the importance of clearly understanding whose perspectives are/were included and whose are/were omitted from a given evaluation. This standard sets the stage for examining cultural perspectives taken in evaluation.

Guidelines. (A) To me, social norms imply a majority viewpoint. I would favor adding or substituting the phrase “cultural norms” to the list of value frames in (A). ◙ A-B There is no one correct approach. The standards should stress analysis from multiple perspectives. ◙ In (B), the complexity of who will make interpretations is not sufficiently visible; e.g., issues of representativeness and who speaks for a stakeholder audience in terms of representing a value position. In (C), add one example to the list of illustrations that makes explicit reference to a culturally congruent strategy. (D) concerns me not for cultural reasons but because it seems to endorse a lack of synthesis in an evaluation report, which is a slippery slope in my opinion. ◙ D. Again the choice may be a combination of options. That is, recognize different value systems by interpreting the data from several perspectives.

Common Errors. I strongly agree with (A). ◙ A. – Amen. ◙In (B), I would add “cultural” to the list of illustrative perspectives in parentheses. ◙ C. Again, not one but multiple criteria may be appropriate to understand the data from several perspectives. ◙Although it would clearly be an error to devote insufficient time to data analysis, I worry that (D) could be used to shortchange time needed for values clarification. Similarly, in (E), I would look for a more appropriate term than “arbitrary” to describe the decision rules of a given stakeholder group. Rules that may appear arbitrary to someone unfamiliar with the culture may in fact have deep cultural significance. Overlooking or failing to educate oneself in such significance should be listed as an additional Error.

Illustrative Case 1—Description. This is one of the few case illustrations that explicitly mentions race, though only urban children of color who scored poorly on standardized tests are addressed. The assumption was made that the source of their difficulty was the fact that they thought and spoke in “nonstandard” English. This is an interesting case to represent values. Though the majority value position concerning Ebonics is never spelled out, presumably it reflected an attempt to respect the children’s cultural expression and start instruction where the children were. Underlying this positive intent, however, is the unwitting condescension that this “nonstandard” English was “less than” Standard English. Parents of Black children differed in their perspectives on whether the alternative curriculum was stigmatizing or enriching, and they raised employability as a criterion beyond test scores. The white parents cited assumed that Ebonics was of lower intellectual content. Teachers were type cast as unwilling to invest the time and energy necessary to gain new competencies. A technical question about the cultural validity of the standardized test norms gets lost in the mix. The discussion becomes political (F2 Political Viability) and the Board aborts the program. Value positions supporting the program are not specifically discussed, nor are the values of the Board members and the Superintendent.

◙ This is one of the most contentious topics in minority education today. Everyone has an opinion and there are multiple perspectives. This topic is so value laden that I first considered it a poor example for illustrating this standard. On the other hand, perhaps it's the best topic. However, one can get caught up in a discussion of the values reflected in the initial decision to change the curriculum, and not give attention to the formative evaluation.

Illustrative Case 1—Analysis. The case analysis does not explicitly address racism, but it does point out the different value perspectives that were relevant to determining the value of this program and hints at the diversity that also exists within a perspective. To me, it would have been more powerful to draw out the different values within stakeholder groups such as Black parents, White parents, teachers and Board. Also, if a bit more information were given about the context of this community, one might be able to illustrate how broader societal values (including prejudices) get drawn into evaluation. The fact that evaluation makes visible basic value differences is an important one. And the fact that evaluation cannot resolve basic value conflicts is also important. However, I think more could have been made of the importance of synthesis, not just asserting that a consensual decision could not be made and leaving it at that. I like the selection of an illustration that has cultural dimensions, but is Ebonics still a timely example? I am not well-informed in this area. If it continues to be used, it should be updated with the best current references and research. Affirmative action, sex education, and drug abuse prevention programs are all fertile examples on which strong values are held.

Illustrative Case 2—Description. Good attention to age diversity and rural location in setting the context for this case, but the term "handicapped" should not be used to refer to persons with disabilities. In the third paragraph, *four* stakeholder audiences are identified: members of the medical research department, the community service organization, the older adult community, and the educational institution. Then something very interesting happens: the values of the medical research department, the community service organization, and the educational institution, and the description proceeds to discuss how these *three* value perspectives were used to shape the evaluation. The values of the older adult community are never mentioned, and in the final paragraph of the description, this constituent group has also disappeared from the decision making structure when "an agreement is reached" about the focus of the evaluation. Given this progression of events, one wonders what the attendance of members of older adult community was like at the second "open" meeting and whether their views were similarly ignored in the selection of success criteria.

◙ There is no mention of what the "older adult community" wanted out of the evaluation. Again the "needs" of the participant group appear to have been ignored. The cultures of the elderly and/or handicapped have unique elements and could have added another perspective/dimension to the evaluation process.

Illustrative Case 2—Analysis. The analyst of this case never notices the disappearance of the older adults from the planning process, instead praising the evaluators for their inclusion of two open planning sessions and for negotiating the values among the (most powerful) stakeholders. ◙ I'm going to rename the participant group, the "invisible stakeholder." ◙I love the potential of this example to address issues of inclusion, engagement versus token representation (Mathie & Greene again here), power and authority—all missed opportunities in this analysis, but a great teaching example nonetheless.

Supporting Documentation. ◙We should definitely find something on age discrimination to accompany this second case example. Also the language piece becomes very salient here again (Rodney's [Hopson] work and Anna's [Madison]) as one examines the language used to communicate values.

Hopson, R. K. (Ed.) (2000). *How and Why Language Matters in Evaluation, New Directions for Evaluation*, No. 86, San Francisco: Jossey-Bass.

Jolin, A., & Moose, C. (1997). Evaluating a domestic violence program in a community policing environment: Research implementation issues. *Crime and Delinquency, 43*, 279-297.

Madison, A. M. (1992). Primary inclusion of culturally diverse minority program participants in the evaluation process. In A. Madison (Ed.), *Minority issues in program evaluation, New Directions for Program Evaluation*, No. 53 (pp. 35-43). San Francisco: Jossey-Bass.

Madison, A. M. (2000). Language in defining social problems and in evaluating social programs. In R. K. Hopson (Ed.) *How and Why Language Matters in Evaluation, New Directions for Evaluation*, No. 86 (pp. 17-28). San Francisco: Jossey-Bass.

Martin, J., & Meesan, W. (2003). Applying ethical standards to research and evaluations involving lesbian, gay, bisexual, and transgender population. *Journal of Gay & Lesbian Social Services, 15*, 181-201.

Pitman, G. (2002). Outside/Insider: The politics of shifting identities in the research process. *Feminism & Psychology, 12*, 282-288.

Presser, L., & VanVoorhis, P. (2002). Values and evaluation: Assessing processes and outcomes of restorative justice programs. *Crime and Delinquency, 48*, 162-188.

Return to Table of Contents

U5 Report Clarity.

Evaluation reports should clearly describe the program being evaluated, including its context, and the purposes, procedures, and findings of the evaluation, so that essential information is provided and easily understood.

Standard. The standard is appropriate at a very general level. Of course, I read "context" as including cultural context, but neither this nor any other dimension of context is spelled out in the general standard. It does bother me that the standard implies that there is a single report and audience. Even at this general level, I believe I would add, "Provided *to* and easily understood *by multiple stakeholder audiences*."

Overview. I believe this overview could benefit from revision to expand attention to non-written communicative strategies and to matters of clarity that extend beyond linguistic translation. Despite a strong opening statement that broadens the definition of "report," the overview focuses the reader's attention on matters of written communication. "Clarity" is given an explicit definition whereas "understandable" is not. By not alluding to any dialog between those doing the reporting and those receiving the report, it positions audiences as passive recipients. Without feedback from audiences, evaluators risk overestimating the extent to which a message is understood (or accepted, which is a whole other matter).

Guidelines. (A) is good, but should be expanded to include cultural considerations that make a reporting mechanism more or less appropriate for a given audience. Either that or a new Guideline could be added that makes it clear that cultural dimensions (history, tradition, audience characteristics, preferred communication styles, rituals, and procedures) should be considered in determining the most appropriate media. (B) is quite specific, and while it is a good fit with many audiences, direct and to-the-point communication is not always culturally appropriate. Perhaps the guideline could be cast a bit more broadly to call attention to dimensions here—length, directness, and scope of focus—rather than stating that brief, simple, and direct are *always* the correct attributes. (C) The idea of tailoring reports to audiences and using multiple media is good and culturally relevant. (D) is expressed in terms of report content, so it should also reference U3 Information Scope and Selection. Cultural context should be named. The more I think about it, the more I favor an additional Guideline to address the cultural context of the report itself and of the reporting process. (E) This one could be expanded a bit to point to "culturally congruent and practice-relevant examples" but the intent of this Guideline is solid; keep it grounded in the real world of the stakeholder audience. (F) I agree with the caution about technical language creating a lack of clarity, but the suggested strategies all appear one-sided (evaluators educate audiences). Opportunities for evaluators to seek advice about clear expression and choice of terms from the audiences themselves (audiences educate evaluators) are not mentioned. (G) It's not entirely clear to me what the referent is here for "problems"—problems of the program or problems of the evaluation. Particularly if it is speaking of additional problems of the program, additional strengths should also be considered (supporting A11 Impartial Reporting). (H) is extremely important to the multicultural validity of reporting (and yes, understandability really is a word; that's one on me!) Important that fairness is included, which should lead to Supporting Documentation citing Ernie House. (I) is good to mention explicitly, though I would separate out oral from written translation so that appropriate methods of forward and back-translation can be cited for written, which would not be used in oral translation. I would favor explicit reference to ASL or other signed language so that the reader is reminded that it is not only a matter of spoken language.

Common Errors. (A) is certainly an error, but framed very much from a position that privileges such "sophistication." I would like to see a parallel concern expressed regarding the cultural sophistication of the evaluator—something like, "Failing to consider cultural variables that define appropriate and effective communication when deciding how to report information." (B) would sit better with me if it said, "Assuming that English is necessarily the appropriate language in which to communicate and that technical terms understood by the evaluators are familiar to the audiences." The error is in not considering these issues, but the current wording puts the non-English speaker in a "one-down" position, as in "taking into account" a personal limitation. (C) "various perceptions" is vague, but it leaves room to consider multiple cultural perspectives. (D) Certainly this is undesirable, but is it setting up a false dichotomy here? Is the implicit message that a report, say, that has been written in the language of the intended audience rather than the technical language of social science necessarily less "precise?" I think "precision" is being encoded as validity here, and I would argue that "clarity" supports validity as well. I fear that the technically infused report is still being privileged regardless of the Guidelines that suggest otherwise. (E) Certainly an important error, albeit one that is only expressed in terms of *written* reports (readability). I would raise a similar concern about the format of an oral or mixed-media report. (F) Not only data aggregation but data synthesis is important to make visible. Through what value frames were the data interpreted in making value judgments (U4 Values Identification; A10 Justified Conclusions)? (G) I'm not quite sure what this is saying—taking up too much time/space describing methods and reporting findings in insufficient detail? Certainly this is not desirable, but again it seems to create a false dichotomy to pit methodology against findings. In order to interpret the findings appropriately, the audience would need to know how the data were collected and analyzed and whose perspectives were included in the conclusions. (H) OK as is, though if one is moving to this level of detail concerning omissions, it might be appropriate to include a statement about failing to take into account the cultural diversity of the consumer population.

Illustrative Case 1—Description. This case describes a 250-page report presented to a School Board, whose members found it difficult to read due to language (technical jargon), format (tests inserted), and level of detail (recommendations lacked clarity due to insufficient detail). No cultural context is given for understanding neither the District nor the background of School Board members or the evaluator, who is internal to this District. The subjects of the report (10 areas) are not spelled out, so it is impossible to intuit what cultural dimensions may have been relevant to consider in relation to these program areas.

Illustrative Case 1—Analysis. The analysis focuses on the format of the written report, with suggestions for improving the format and augmenting the full written document with additional communications. The suggestions all appear plausible for a School Board as the audience, but it would be more helpful if the analyst had made visible the ways in which the Board environment shaped his/her format suggestions. In other words, with only one audience addressed, one loses the sense in which the report would be different for a different audience.

Illustrative Case 2—Description. The context here is business rather than K-12 education; the evaluand is a new training program developed by an instructional design team in a business setting. Here again, no other context information is provided nor is the topic of the training program known, without which it is impossible to determine relevant cultural considerations. As in the first Illustrative Case, the emphasis is on the length and format of the report. Presumably the "technical jargon" referred to is evaluation jargon, because if it were jargon related to the business setting or to the specialized area of training addressed by the program, it would not be inappropriate. The reference to the format limitations of word processing software is certainly dated, though considerations of visual attractiveness are still relevant.

Illustrative Case 2—Analysis. The analyst makes detailed recommendations in the first paragraph for reordering sections of the report so that busy professionals can find key information quickly. There is clearer reference here to the busy world of professionals and to the fact that written material (especially lengthy written material) may not be the most effective communication strategy. The idea of a multimedia presentation using visual displays and graphs seems congruent with instructional design in a business environment, though the point is not made that the analyst is seeking to match communicative strategies to the (organizational) culture of the setting. These cases are both skeletal and similar. Neither of them brings out cultural dimensions of report clarity well.

Supporting Documentation.

Burker, J., Minassians, H., & Yang, P. (2002). State performance reporting indicators: What do they indicate? *Planning for Higher Education, 31*, 15-29.

Hopson, R. K. (Ed.) (2000). *How and Why Language Matters in Evaluation, New Directions for Evaluation*, No. 86. San Francisco: Jossey-Bass.

MacNeil, C. (2000). The prose and cons of poetic representation in evaluation reporting. *American Journal of Evaluation, 21*, 359-367.

Madison, A. M. (2000). Language in defining social problems and in evaluating social programs. In R. K. Hopson (Ed.), *How and why language matters in evaluation, New Directions for Evaluation,* No. 86 (pp. 17-28). San Francisco: Jossey-Bass.

Stockdill, S. H., Duhon-Sells, R. M., Olson, R. A., & Patton, M. Q. (1992). Voices in the design and evaluation of a multicultural education program: A developmental approach. In A. Madison (Ed.), *Minority issues in program evaluation, New Directions for Program Evaluation*, No. 53 (pp. 17-33). San Francisco: Jossey-Bass.

◙ I would add something by House—not sure what citation to use here, but something on fairness. Also, we could update the literature on reporting strategies, modes of presentation, etc. Especially examples that include cultural content.

Return to Table of Contents

U6 Report Timeliness and Dissemination.

Significant interim findings and evaluation reports should be disseminated to intended users, so that they can be used in a timely fashion.

Standard. This standard seems unnecessarily restrictive in terms of the dimensions of time it addresses, the focus on results-based issues of time to the exclusion of process-based issues, and the attention to intended users as opposed to broader audiences. Time is a matter of great relevance to cultural competence in a number of ways. This standard, as written, does not do it justice. I would actually favor a separate standard on matters related to time and timing, apart from the dissemination issues that this standard takes up. I would agree that a standard on dissemination issues is still important to retain, however.

Overview. The overview first alludes to the importance of getting the information to an audience when the information can best be used—hence the emphasis on timing—but this point is not well developed. As already mentioned, I think matters of time frame and timing are deserving of their own standard. Here, they get lost in the mix. The overview does better in addressing the issue of entitlement; specifically, who is entitled to see the results of the evaluation? (their definition of intended user). This is an important and interesting discussion. I like the inclusion in (item 4) of those who provided information to the study as a group entitled to receive results so they can see how their data were analyzed and interpreted. I also support (in item 5) a broad definition of stakeholder audiences, though the justification for someone being a stakeholder is not well presented, making the listing of potential stakeholder categories—parents, students, media—less than useful. Third is the issue of format or communicative strategy, cross-listed with U5 Report Clarity. The need to tailor a report to fit cultural practices is acknowledged, along with the potential need for language translation, but again these issues are not further developed. The final paragraph of the overview expands on important matters of power and authority, responsibility and control over the dissemination process, focusing on the evaluator-client relationship and compliance with or reasons to override Formal Agreement (P2). This is an extremely important, culturally-relevant discussion, and it could be illustrated with a case that draws out dynamics of power better than the cases provided. Overall, I think this standard tries to encompass too much. It does a good job of addressing issues of authority and entitlement surrounding dissemination. It is less successful in the areas of timing and actual strategies. I think it would be more effective if divided into more than one standard.

Guidelines. (A) These rational, linear "ground rules" reflect a majority perspective and may themselves be culturally incongruent. Stakeholder inclusion is good but should not be entered into with a fixed agenda of what information is relevant to timing and dissemination issues. These questions *assume* that a report is desired, for example, whereas Patton would start off questioning that assumption. In referring to "representatives of the key stakeholder groups," it begs the question of how such persons are identified or selected, a matter deserving of explicit attention under U1Stakeholder Identification. (B) *If* one is operating with the presumption of a report and in a linear, monochronic time frame (Ing, 2001), these are reasonable guidelines. I suspect that these procedures are culturally bound in ways that are not explored. (C) Here again, *if* one is using a preordinate design in which these things are known at the outset, this is a good fit. Less so if one's design is emergent. (D) Again, OK *if* assuming a linear timeframe and intended users known at the outset. I guess I would favor guidelines that directed one to explore understandings of time frame, the notion of "deadlines," and the expectations for what the process of sharing information might look like (the "report" piece). (E) Interesting assumptions of control over the process here. Whose schedule? This guideline would be a nice fit with many evaluation contexts but be incongruent with others. Assumes a preordinate design that includes a fixed schedule. (F) This meta-evaluation guideline should cross-reference A12. While not explained in this way, it is potentially a very relevant tool to enhance cultural competence. Key details, of course, would hinge upon who is considered a "qualified person" (shades of U2 Evaluator Credibility) and how "quality" is defined (U4 Values Identification). (G) If such agreements are culturally congruent, the language in this guideline would be appropriate, but in some contexts (and with some communicative strategies) issues of "editorial control" and "intermediate and final reports" as envisioned here may not be relevant. (H) Could definitely be used to support cultural competence. What else besides clarity and factual accuracy would be important to note? As in (A), it does not address how appropriate representatives might be identified. Also the notion of fixed time frame again here, against which some release of findings, could be judged to be "premature." (I) The idea is good, but the examples traditional. Push the envelope here to help evaluators think outside the box regarding communicative strategies. (J) **This one's a zinger—diversity framed as a "social impediment!"** (J) Definitely needs to be rethought and reworded. Instead of casting diversity as a barrier or impediment, it should speak to drawing upon the strengths of cultural traditions and practices in identifying the most appropriate communicative strategies and timing information exchange.

Common Errors. (A) has potentially important implications for cultural competence, though it does not address the issue of how to move a constituency into consideration as an intended user if the intention is to exclude. (G) touches on similar issues but from the benign stance that stakeholders "who do not have spokespersons" might (inadvertently?) be ignored as opposed to deliberately excluded. Again, dynamics of power and privilege here; how does one gain a spokesperson, gain a voice? (B) should be cross-referenced to A7 Systematic Information Control. (C) and (D) are based upon assumptions of time as linear and monochronic, as already discussed. (E) is actually a very significant point that almost gets buried here, I think. It speaks to Scriven's notion of "overrides" when synthesizing evaluation data (cf. Scriven's Key Evaluation Checklist, under Significance) –considerations that would interrupt or override previously specified plans for data synthesis. The key issue for us would be to get civil rights, social justice, and equity issues seen as potential overrides that would be viewed in the same manner and with the same response as the violations listed. (F) While I'm not supporting any violation of the Accuracy standards, I do think it's interesting to note the hidden assumption here is that a report reaches a stage at which its content is both accurate and complete. To me, this links to validity (A5) and I've always been partial to Cronbach's cautionary note that our understandings are *always* partial and incomplete, pending the results of the next study (not a direct quote but the gist of his sentiment).

Illustrative Case 1—Description. This is a clear-cut case in which an evaluator fell behind schedule and apparently cut corners by not following a phased dissemination of findings to relevant stakeholders. Other than the fact that the setting is a school, no context information is provided at all.

Illustrative Case 1—Analysis. Reasonable observations, given what little is known about the setting, but an amazing number of "*should*" gives the impression that there is only one right way to move through this scenario, and that the correct path is easily discerned by an outside observer (the author of the analysis). I'd prefer to see an analysis raise questions and suggest possibilities without implying such a singular view of what compliance with the standard looks like. But this is just me. What do the rest of you think about the tone of the analysis?

Illustrative Case 2—Description. This case description, set in a higher education context of a College of Nursing and Allied Health Professions, is odd in that the description of the evaluation provides no evidence to support the decision made by the Board of Trustees. The bulk of the case description focuses on the methods used to calculate and present attrition rates, whereas the Board concludes that there are quality problems with the programs. This case seems to be a clearer violation of Justified Conclusions (A10) than of U6. Of cultural relevance may be the fact that no mention is made of diversity characteristics to see if there were differential attrition among subgroups, and one could also examine the variables included in the attrition formula, only some of which—SATs, GPAs, and age at admission—are identified in the scenario.

Illustrative Case 2—Analysis. Given the organizational culture of an institution of higher education, the suggestions in the analysis seem viable. I like the slightly greater complexity of this case insofar as doing something correct—providing an executive summary—backfired because of a different problem—not getting the full report out soon enough. Interesting, however, that the analyst assumes timing was the only issue in the failure to read the report. As described, the report may have been unclear (U5 Report Clarity) at best and even given more time, Board members may have focused on the Executive Summary. The links to U3 Information Scope and Selection and A9 Analysis of Qualitative Information also seem appropriate.

Supporting Documentation.

Ing, C. (2001). Culturally appropriate evaluations. In P. A. Gabor, R. M. Grinnell & Y. A. Unrau (Eds.), *Evaluation in the human services* (pp. 285-303). Baltimore, MD: Peacock.

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

Winberg, A. (1991). Maximizing the contribution of internal evaluation units. *Evaluation and Program Planning, 14*, 167-172.

Return to Table of Contents

U7 Evaluation Impact.

Evaluations should be planned, conducted, and reported in ways that encourage follow-through by stakeholders, so that the likelihood that the evaluation will be used is increased.

Standard. This is an extremely important standard. While not framed in terms of cultural competence, it in fact could be read as a mandate for such, given that cultural competence is understood to maximize the likelihood of evaluation use (a hypothesis that has yet to be tested?)

Overview. What is most noticeable about this standard is that it is grounded in a traditional definition of use that is exclusively results-based. It should be updated to reflect broader constructions of evaluation influence (Kirkhart, 2000). Within a results-based framework, it defines impact as inclusive of instrumental and conceptual use but it does not encompass persuasive use. It also assumes that impact is positive, failing to consider unintended influences of evaluation or intended influences that may be experienced as negative from the vantage point of certain stakeholders. The time frame(s) in which impact occurs is not addressed. The evaluator’s role in facilitating use is addressed, and the point about use not being automatic is certainly well supported by literature; however, the way in which it is written to me sounds condescending of program persons. The evaluator is portrayed as a helper who can show them the way,

rather than as a collaborator or consultant who can work with the program persons and stakeholders to explore different options.
Guidelines. (A) continues to portray the evaluator as someone who knows the right answer and the stakeholders as persons to be educated or convinced. I'd prefer language that directed the evaluator to work with stakeholders to identify ways in which the evaluation findings might be useful for their work. Then, of course, I'd continue the conversation to discuss ways in which the process of planning and carrying out the evaluation could be useful. (B) is good, but I would probably say "participating in" rather than "assisting with" so that the power differential between evaluator and stakeholder is not built into the standard itself. (C) is headed in the right direction, but with the caveat that "open, frank, and concrete" may not be the defining characteristics of culturally competent communication in a given context. Also, to me, "concrete" implies, "You'll have to break it down, because stakeholders may be too stupid to understand." Appropriate cross-referencing of U5 Report Clarity and P6 Disclosure of Findings. (D) is appropriate within the conceptual constraints already discussed in U6 Report Timeliness and Dissemination, which is well cited here. (E) is well written; to me it communicates greater respect for stakeholders insofar as there is no implied power differential with evaluators. (F) is good in that it acknowledges the value of using multiple communicative strategies but limited in that it alludes only to written and oral communication. The broader message here should be that one must attend to cultural context in determining the mix of communicative strategies that will be appropriate and effective. (G) again assumes the time frame is linear, the influence results-based, and the impact unidirectional (evaluator helping stakeholders). It should be rewritten to broaden the conceptualization of ways in which evaluators can work with stakeholders to support the impact of their work.
Common Errors. (A) correctly notes that it is an error to communicate disrespect for stakeholders; yet as (E) points out, issues of influence are complex. Clients and stakeholders may hold perspectives, values and worldviews that are very different from those of the evaluators. (C) is an important caution, citing U4 Values Identification, though (B) to me suggests a false dichotomy between theory and practice. The language of (G) should be rethought to eliminate the word "target" which projects all sorts of power issues as well as safety vs. harm; the idea of maximizing impact by attending to the needs of specific stakeholder audiences is fine. (F) and (I) both take up issues of *mis*use of evaluation, and extremely important issue that gets a little buried. Perhaps this standard needs to take up both sides of the issue—facilitating appropriate use and guarding against misuse—in a more balanced way or perhaps it should be split into two standards. This opens up a complicated discussion that circles back to theory of use. (D) and (H) speak to the roles of client vs. evaluator and connect with conversations on evaluator role and evaluation recommendations (Cf. Scriven's discussion of Recommendations in the Key evaluation Checklist). These conversations are certainly culturally bound, often in terms of organizational or community culture rather than personal demographics.
Illustrative Case 1—Description. This is an interesting case in that it reflects the complexity of matters surrounding impact, and it directly (though not explicitly) introduces issues of power and authority. The context is described as an elementary school, the subject matter reading, and the audience parents of the elementary school

students. The evaluator is a woman, and she is officially designated as the "reading specialist" of the district. This is an example in which more detail on the educational level, economic background, gender, age and race/ethnicity of the key players would be helpful in drawing out some of the reasons behind the parents deferring to the evaluator and withdrawing from the project. It also speaks to the importance of the evaluation remaining congruent with the values of the program being evaluated. Since the program intent was parent participation, this theme could have fruitfully been carried over into the evaluation. A "top down" evaluation is incongruent with a participatory program. This indeed illustrates evaluation influence; the evaluation undid what the program was trying to accomplish! Good that a case illustration of negative influence is provided, though the case analysis itself does not explicitly make this point.

Illustrative Case 1—Analysis. Because no cultural information is introduced in the case itself, the analysis cannot explore the possible influences of similarities/differences in age, gender, education, economic resources, and race/ethnicity among the parent group and between the evaluator and parents. Lacking such details, the analyst is left with the rather superficial observation that the evaluator was "overzealous" in extending herself to provide suggested revisions. Issues of power and authority are not addressed. The parents had called her in as an "expert" from the outset, a role definition that could have been renegotiated early on to improve congruence of the evaluation and the program. The analysis does not take up this broader issue of ideological congruence between evaluation and evaluand.

Illustrative Case 2—Description. I like the inclusion of a positive example so that readers can see what the authors view as compliance with a standard; most of the cases illustrate violations of standards. Also, this is set in an industrial context, though again additional context information is missing. This case is written at such a general level that it's hard to garner much of anything from it. To be meaningful, one would need to have some sense of the "physical and/or verbal behaviors" of relevance that were being observed. Even more salient to our cultural reading would be the specification of what "trainee characteristics" were examined to determine the reasons for lack of progress (p. 62, last paragraph of the description).

Illustrative Case 2—Analysis. This evaluation is deemed successful because the evaluator and the trainer spoke the same language and shared the same goals. The perspectives of the trainees themselves are not represented in the illustrative case nor queried in the analysis. Since this case is put forth to illustrate Evaluation Impact, impact on the consumers should also be considered. The analyst incorrectly asserts that "appropriate stakeholders" were encouraged to follow through, without commenting on the trainees.

Supporting Documentation.

Aver, L., & VanTassel-Baska, J. (2001). Investigating the impact of gifted education evaluation at state and local levels: Problems with traction. *Journal for the Education of the Gifted, 25*, 153-176.

Cherin, D., & Meezan, W. (1998). Evaluation as a means of organizational learning. *Administration in Social Work, 22*, 1-21.

Huba, G., Brown, V., Melchior, L., Hughes, C., & Panter, A. (2000). Conceptual issues in implementing and using evaluation in the 'real world' setting of a community-based organization for HIV/AIDS services. *Drugs and Society, (16)*, 31-54.

Kirkhart, K.E. (2000). Reconceptualizing evaluation use: An integrated theory of influence. In V. J. Caracelli & H. Preskill (Eds.), *The expanding scope of evaluation use, New Directions for Evaluation*, No. 88, (pp. 5-23). San Francisco: Jossey-Bass.

Yeh, S. (2000). Building knowledge base for improving educational and social programs through planned variation evaluations. *American Journal of Evaluation, 21*, 27-40.

Feasibility Standards

The feasibility standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal.

F1 Practical Procedures.

The evaluation procedures should be practical, to keep disruption to a minimum while needed information is obtained.

Standard. This is an interesting standard. Given the labor-intensiveness of culturally competent practice such as primary of inclusion of consumers in the evaluation process or the initial rapport-building with key stakeholder audiences, one might find those procedures challenged as impractical, citing this standard. However, this standard could also be used to support a respectful interaction between evaluators and program providers, noting that the evaluators should take care not to disrupt program operation more than absolutely necessary. The presence of evaluators will always be somewhat disruptive.

Overview. The definition of procedures is sound, but the list—though not intended to be inclusive—omits any mention of context or voice. Minimally, the list should include "determining what dimensions of cultural context are most salient" and "identifying key stakeholders." The distinction in the last short paragraph between "theoretically sound" and "unworkable" bothers me in a way that I can't quite pinpoint. I agree with the basic message that textbook perfect evaluation designs must consider the workings of real-world settings and that the two may clash, but doesn't validity demand that it be *both* theoretically sound *and* workable? I think the way it's written gives the impression that one could have a perfectly good evaluation if only the program weren't in the way, whereas by definition good (valid) evaluation *must* include the program.

◙ Overall, this section does not mention that the diversity of the evaluation population may impact the procedures listed. For example, in the first paragraph of the overview, add a statement to the effect, "The procedures listed above should be undertaken in a manner that considers the diversity of the population served and the stakeholders."

Guidelines. (A) is a key opening to introduce cultural competence as a dimension of "qualified personnel." I'm not sure that "training" quite captures the remediation needed to redress a lack of qualification in this area, though certainly workshops, etc. are not an irrelevant strategy. ◙ A. add a sentence: "Personnel should be culturally sensitive as well as trained in evaluation techniques in order to address the characteristics of diverse populations." ◙ (B) is an appropriate concern. Perhaps the issue of how much effort is "reasonable"—e.g., to enhance multicultural validity—is broader than this standard. (C) Clearly there are time constraints and people constraints on all evaluation. My concern is that key audiences may be omitted because it is just too time-consuming to work with them (or evaluators may lack the skills to do so). Does cultural competence sometimes demand "impractical" procedures (at least when viewed from a majority perspective)? I'm guessing that it may, and somehow this needs to be calculated into the mix. (D) seems fine, though the cultural congruence of evaluation with "routine events" should be questioned. (E) is a reasonable guideline for a preordinate design; is this principle already built into emergent designs? (F) suggests good procedural checks, but they should include stakeholders beyond the client. The client may not be aware of issues of timing and availability from the perspectives of all relevant groups. (G) Pilot testing is good for many reasons, one of which is the practical matter of timing addressed here. ◙ G. add “, and whenever possible, those taking the pilot test should represent the diversity of the population.”

Common Errors. (A) is a good recognition of the importance of setting. I would favor adding “or cultural context” to cue the reader to potentially broader dimensions that inform “fit.” I like the tie to validity in (B), with the recognition that an evaluation lacking, for example, access to certain perspectives may not be useful. Some “practical” constraints may be sufficiently serious as to invalidate the study, and it is best not attempted. I agree that (C) is undesirable, and I would add (D) Failing to consider cultural competence in selecting evaluation personnel qualified to craft an evaluation that is congruent with context.

Illustrative Case 1—Description. The context of this illustrative case is elementary education, in schools with “high concentrations of economically disadvantaged students” varying in location among urban, suburban, and rural. The evaluators selected a randomized control group design over three years, not considering the likelihood of attrition as well as movement of students among the sites over this period. The control group designation was also compromised when the experimental intervention (funding) was equalized among groups by administrative actions. No other details of cultural context are provided beyond location and economic status.

Illustrative Case 1—Analysis. This analysis correctly points to the complexity and the politics surrounding educational systems and their interventions. Though student attrition and mobility could have cultural origins (e.g., migrant populations), this is not brought out in the case description, making it difficult to speculate on such in the analysis. ◙ Case 1: Clearly the evaluator was unaware of the “culture of poverty” including frequent family relocation. This could be referenced as such in the analysis.

Illustrative Case 2—Description. This case, set in a business environment, discusses a comparative evaluation of two alternative training models in computer programming. No information is given on cultural context, including organizational culture, but the description implies that the steps taken by the evaluators were congruent and well-received. There is good detail in this case description, but it is entirely on procedural/methodological matters of the design and its implementation. It doesn’t actually address the extent to which disruption was minimized, so the connection to F1 is a little indirect. Also, the content is dated, given advances in computer programming and on-line instruction since this was written. ◙ Case 2: Some of the steps taken, although not specifically mentioned, must have focused on gaining an understanding of the “culture of the organization”, including participant views on learning computer programming. And the evaluation had to be responsive to that. As currently portrayed, the case ignores these factors, thus implying attention to them is not needed.

Illustrative Case 2—Analysis. The analysis focuses on the careful preparation and backup plans for the study as well as the time required to communicate with participants and “representatives of different interest groups.” The point regarding how much time it takes is extremely important. It should be drawn out in more detail in the description to support the analyst’s discussion. Also the description does not make clear the perspectives of the various interest groups nor how issues of power and authority were addressed when instructors, managers, participants, and observers came together in the focus group. This case should balance design details with details concerning the communicative processes to strengthen the connection to F1.

Supporting Documentation.

Alkon, A., Tschann, J., Ruane, S., Wolff, M., & Hittner, A. (2001). A violence prevention project with ethnically diverse population. *American Journal of Preventative Medicine, 20*, 48-55.

Huberman, M. (1996). A critical perspective on the use of templates as evaluation tools. In M. A. Scheirer (Ed.), *A user's guide to program templates: A new tool for evaluating program content*, *New Directions for Program Evaluation* , No. 72, (pp. 99-108). San Francisco: Jossey-Bass.

◙ None immediately comes to mind. We need something on routinizing evaluation or on the cultures of practice and evaluation. Ideas?

Return to Table of Contents

F2 Political Viability.

The evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained, and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted.

Standard. This is an extremely important standard. The "various interest groups" holding different positions may break along cultural lines, though this is not explicitly addressed. The standard addresses the general issue from both a positive (facilitating cooperation) and negative (preventing bias and misuse) perspective, although the language seems to privilege the evaluator role, leaving the interest groups to either "cooperate" or "curtail."

Overview. Interest groups are defined as any group seeking to influence policy toward a shared goal. Issues of power are explicitly addressed in this standard; evaluation controls influence and resources. What is not explicitly taken up in this standard is the fact that evaluation itself may be seeking to influence policy toward a shared goal; advocacy models of evaluation are obvious illustrations. Fairness and equity are introduced in a very limited way in the definition of political viability: "Evaluations are politically viable to the extent that their purposes can be achieved with fair and equitable acknowledgement of the pressures and actions applied by various interest groups with a stake in the evaluation." (p. 71) I'm not sure it's so much about *acknowledging* such pressures and actions as it is *balancing* or controlling them fairly and equitably. The overview seems to soften the standard; it lacks teeth. The second paragraph of the Overview returns to the positive/negative theme of the standard itself, both cautioning and commending evaluators regarding political pitfalls and victories, respectively. It seems well balanced to me. ◙ Or is it an example of dichotomous thinking? (cf. Patricia Hill Collins) ◙ This standard could be revised to better address cultural diversity.

Guidelines. (A) seems appropriate irrespective of the judged volatility. All evaluations are "potentially volatile." Also, the context of power and authority within which evaluation is often conducted seems a little understated here. Evaluators must be careful not to offer false reassurances or to promise more than they can deliver. ◙ Guideline A could include a qualifier, “Interest groups from diverse backgrounds may not see the need for evaluation, and may resist the process entirely. Evaluators should be prepared to assuage these concerns and engage all relevant stakeholders.” ◙ (B) for a preordinate design, the level of detail in the contract is appropriate. From a diversity perspective, the contract should include review and amendment as necessary to maintain congruence between evaluation methods and cultural context. (C) involves key audiences, which could include culturally diverse stakeholders, but their role is portrayed passively. “Not being surprised” by outcomes is quite different from being included in the process of data synthesis and data interpretation. (D) can support cultural competence, but given how labor-intensive primary inclusion can be, I would prefer to reframe this as, “Budget adequate resources to support the inclusion of different perspectives.” Otherwise, I fear that resource constraints could be used as a ready justification for exclusion. (E) appropriately introduces the public’s right to know as an ethical principle, congruent with the AEA *Guiding Principles*, and acknowledges the possibility of evaluation doing harm to the extent that terminating the study may be the most responsible course of action. To this list of guidelines, I would add (F) Make explicit the stakeholder perspectives that were represented in the study and those that were omitted, acknowledging than any single study will necessarily have limitations of perspective (see A10 Justified Conclusions).

Common Errors. (A) may be more than a matter of appearance. Errors could be real *or* apparent imbalances. (B) addresses organizational power structure, but similar mention should be made of societal dynamics of power and privilege, many of which are associated with cultural diversity. (C) seems all right as written, but (D) is too cryptic. First of all, the issue of “objectivity”—even though placed in quotes in the text—breaks along epistemological lines that should be acknowledged. The point being made is an extremely important one; fairness in evaluation is an important standard that does not reside in any single method. Rather, like validity, it resides in the application of methodologies. It seems to me that the Common Error would be to assume the fairness of any given methodology without consideration of the consequences of its use. ◙ Error D could include, “ ... This assumption may significantly impair your evaluation, specifically biasing it against diverse communities.”

Illustrative Case 1—Description. This case is set in a K-12 public school district with the main characters being a team of professors serving as evaluators, the District Superintendent, and the teachers’ union. The evaluand is a state-funded “innovative approach to reading instruction,” and the evaluation is mandated by the state. No information on cultural context is given. The Superintendent allegedly opposes the program because it limits his flexibility in allocating district funds, though the scenario does not state what alternate use of funds he values so much that he is willing to lose the state contribution. The lack of detail makes it difficult to discern his motivation. Also his relationship with teachers and the teachers union prior to this evaluation is not explored. The disputed design component is the assessment of teachers’ instructional skills, but the illustration lacks sufficient detail to appreciate the relevance of this design component to the evaluation questions posed by the state. For example, if this were part of an intervention check to assure that the “innovative approaches” were actually implemented, it would be more relevant than if it were a thinly veiled personnel evaluation. The language of the scenario subtly impugns the professionalism of the professors by referring to the teacher assessment component as a “scheme.” Clearly the professors erred in their Stakeholder Identification (U1) by not dealing with the teachers’ union up front. Consistent with the scenario, the Superintendent’s opposition to the evaluation is hardly a surprise; his agenda is to remove the program and regain control over budget allocations irrespective of the evaluators’ actions. A key missing point is how the evaluation was received at the state level and by the teachers’ union and the teachers themselves. Also, no mention is made of either student or parent perspectives on this program.

Illustrative Case 1—Analysis. The analysis minimizes the political agenda here, asserting that if the Superintendent had agreed in writing to the evaluation design, it would have precluded dismissal of findings. In fact, the Superintendent would likely have found another basis of objection, since this is a political matter, not a scientific one. There are broader issues of power and control between the Superintendent and the state. The case offers no clue as to what evaluation guidelines came with the state mandate. The analysis correctly faults the evaluators for omitting the teachers’ union from early conversations, but again ignores the broader historic contexts of ongoing tensions between this Superintendent and the union as well as the union’s stand on other evaluations in the past. The advice of the analyst to “deal with” tensions before the study was conducted incorrectly implies than an evaluation can be stripped of its political context through early procedural steps. In fact, the overarching error of the professors appears to have been the fact that they did not understand this study to be a political animal as much as it was an empirical one.

Illustrative Case 2—Description. The evaluand in this case is a collection of 14 remedial education and job training programs for “economically disadvantaged youth, ” introducing age and economic status as relevant diversity variables. No other cultural dimensions are mentioned nor is the influence of age and economic status further explored. The politics surrounding these programs is not explicitly addressed, though the possibility is raised of a covert agenda to use the evaluation as a means to reduce funding. The programs are federally funded but administered by a state agency, and the evaluation is mandated. Presumably both the state agency personnel and the legislators have expressed interest in reducing the funding of youth programs as a response to budget crisis. The evaluators are asked to complete the study in three months, a time frame that they recognized as insufficient to complete a definitive outcome study suitable for informing refunding decisions. Evaluators are “in-house—presumably at the state agency level. Issues of evaluator credibility (U2) are raised, but there is no mention of how they were dealt with. The case focuses on the actions taken by the evaluators to deal effectively with a political standoff between the program operators and the state agency. Evaluators renegotiated the contract with the state agency to make clear what could reasonably be accomplished in the allotted time frame, and program administrators were well included in the evaluation process and reporting. Interestingly, no other stakeholder perspectives—e.g., the youth participating in the programs or members of their families or communities—are mentioned. Also, there is no mention of whether legislators ultimately reduced funding for the program on bases other than the evaluation.

Illustrative Case 2—Analysis. The analysis focuses on the steps taken by the in-house evaluators to mediate or eliminate conflict and antagonism, supporting the “utility and mutual benefit of the evaluation to both the program operators and the agency administrators." The larger political context of the programs and their evaluation is not scrutinized, however, leaving the reader with a false impression that conflict is necessarily a negative aspect of evaluation and that it is avoidable if one follows certain procedural steps.

Supporting Documentation.

Chelimsky, E. (1995). The political environment of evaluation and what it means for the development of the field. *Evaluation practice*, *16*, 215-225.

House, E. R. (1993). *Professional evaluation: Social impact and political consequences*. Newbury Park, CA: Sage.

Return to Table of Contents

F3 Cost Effectiveness.

The evaluation should be efficient and produce information of sufficient value, so that the resources expended can be justified.

Standard. A complex standard, laid out in deceptively simple terms. Much hinges on whose values define costs and benefits and who incurs each. Also, the equation omits time frame, and much hinges on the point in time at which the equation is calculated (e.g., how long will one wait to examine benefits? History may put the opportunity costs of the evaluation in a different light.) The construct itself is culturally bound; not all contexts may find such a rational, industrial metaphor a good fit for how evaluation is viewed.

Overview. The initial definition of costs includes money and non-money costs, but not opportunity costs (either money or non-money). It seems to me that the costs of an evaluation cannot accurately be calculated if the opportunity costs are omitted. (For example, in considering the costs of my School's self-study for reaffirmation of accreditation, one would have to consider not only the financial expenditures and the countless hours spent, but also the value of what could have been done with that money and especially time, had it not been spent on accreditation.) The definition of benefits is exclusively results-based and therefore insufficient. Benefits of the process of evaluation should also be considered and illustrated (e.g., values clarification, empowerment of certain stakeholders, opening new lines of communication).

The second paragraph of the Overview makes the entire process of deciding whether or not to conduct an evaluation sound cut-and-dry, as if it is a simple matter of running the numbers. It assumes a preordinate stance; i.e., that all of the relevant costs and benefits can be known in advance. This framing can endanger costly, labor-intensive procedures required to establish multicultural validity in certain contexts. An affordable design that is invalid is a questionable investment.

The third paragraph of the Overview introduces some of the complexities of determining cost-effectiveness, though these are understated. The differential weightings of outcomes may break along cultural lines as different stakeholder interests are more or less considered. Advising evaluators to rely on past experience and seek second opinions is insufficient. More explicit consideration should be given to broadening the elements and perspectives included in cost-effectiveness. This paragraph should appear second in the Overview, with expanded discussion of the complexities of this framing. Prudence and efficiency, cited in the last paragraph, are more matters of Fiscal Responsibility (P8) than F3.

◙ It should be noted that several other standards qualify the level of stakeholder involvement as a function of monetary and time constraints. This could be addressed with an additional guideline and error.

Guidelines. (A) Although the costs of materials and services are perhaps the simplest to identify, that is a reasonable starting place. Software and other technology-related costs should be included in the examples. While a budget (B) is clearly an appropriate planning tool, a Gantt Chart or other timeline tool is more likely to make visible the non-money costs of the evaluation process and who will incur them. (C) seems more relevant to F1 Practical Procedures than to F3. The implication appears to be that disruptions drive up costs, but unexpected disruptions to preordinate plans can also yield benefits. I would drop C. An inventory of benefits (D) is also a relevant planning tool, but the phrasing of D privileges the benefits to the client over those of other stakeholders (whose benefits are referred to as side effects). The Guidelines should address both the identification of benefits and the determination of their relative importance, but as separate planning steps. (E) is all right, but with the caveat that this equation may shift as the evaluation is underway and it should be revisited during the evaluation process. (F) belongs under P6, as referenced. To include it here minimizes the complexity of cost-effectiveness, which does not always reduce to matters of economy. I would delete F. (G) Although "proper balance" is not defined, the intent is good. This is a matter of opportunity costs that needs to be more explicitly addressed—the costs of the evaluation viewed in light of the value of what the program could accomplish with those same resources. This is definitely part of the equation of cost-effectiveness. ◙ An additional guideline, "Stakeholder involvement in the evaluation should be considered as a necessary cost in developing a budget."

Common Errors. (A) is an error, but it should be explained that even though costs are fixed, the benefits can be maximized by full consideration of stakeholder perspectives. (B) would be an error even without considerations of cost-effectiveness. The selection of method is grounded in many more standards than this one. (C) is clearly an error, but extends beyond cost-effectiveness. Perhaps a separate standard is needed to address the adequacy of resources allocated to support evaluation. (D) Same comment as (A)—the "error" warrants explanation regarding the delineation of benefits. (E) I strongly support this point and suggest that it be moved up in the list to a position of greater prominence. Culturally competent procedures may be labor-intensive and time-consuming, yet enhance the validity of the evaluation. This could be cited as an example or brought out through an illustrative case. (F) reiterates points of the overview that are important to highlight; I support the redundancy. The point regarding differential value allotted by different groups deserves its own Guideline/Error so that it is not lost in a more generic statement of the complexity of cost-effectiveness calculations. What is valued by one cultural group may not be recognized as valuable by another (see U4 Values Identification). ◙ An additional error, "While stakeholder involvement is necessary, their level of participation should not exceed the cost effectiveness of their involvement."

Illustrative Case 1—Description. Given the complexity of this standard, this is a disappointingly simplistic case. To evaluate an elementary school math program, evaluators tested third and fourth graders on 300 computational problems, taking a total of 7 ½ hours to complete. No other context information is given. This goes beyond cost-effectiveness to sheer incompetence on the part of the evaluator in my opinion (see U2 Evaluator Credibility). This case illustration should be replaced with one that can bring greater complexity to the conceptualization of costs and benefits from multiple perspectives.

Illustrative Case 1—Analysis. The analyst is more constrained than I might have been given such an outrageous design on the part of evaluators. He/she correctly outlines a more appropriate sampling plan that produces a more cost-effective means of generating district-level scores. Following the lead of the case, the analysis is written strictly as a matter of methodology and costs, missing larger issues of values and of benefits of this evaluation.

Illustrative Case 2—Description. The setting for this case is a medical school, and the case takes as given the fact that "policy requires that student achievement in third-year clerkships must be evaluated, in part, by the use of 'objective measures of acquired knowledge.'" The remainder of the case describes the costs of purchasing external, nationally standardized exams, with some data provided on their reliability and validity. There is no mention of the opportunity costs of purchasing the exams in terms of either program expenditures or the other evaluative methods that are presumably used if these objective measures constitute only part of the evaluation of student performance. This evaluation is mandated by policy, and any benefits beyond fulfilling the mandate are not explored. The directors of the five clerkships selected the exams; no other stakeholder values are discussed.

Illustrative Case 2—Analysis. The analysis is quite simplistic. Noting a criterion-related validity coefficient "in the low .80s" between the standardized tests and locally developed tests, the analyst asserts that it may be more cost-effective to develop the tests locally. Astonishingly, the opportunity costs of faculty spending their time on test development, validation, and norming are summarily dismissed: "Despite the convenience and faculty time savings, the directors might find that the local preparation would be less expensive." The analysis does make reference to the potential benefits (termed positive side effects) of local test development (refinement of the clerkship content), which is good, though the benefits of the nationally standardized tests are not similarly addressed.

Neither of the illustrative cases addresses the cultural dimensions underlying the calculation of costs and benefits. While the Overview, Guidelines, and Common Errors allude to at least some of the complexity of this checkpoint, the case illustrations are underdeveloped and simplistic and do not support a full understanding of this standard.

Supporting Documentation.

Levin, H. M., & McEwan, P. J. (2001). *Cost effectiveness analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Watson, D. (2002). Just a paper exercise. *Social Work Education, 21*(1), 79-89.

Return to Table of Contents

Propriety Standards

The propriety standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.

P1 Service Orientation.

Evaluations should be designed to assist organizations to address and effectively serve the needs of the full range of targeted participants.

Standard. This is one of my personal favorites among the standards. Added in the last revision, it was also much debated, challenged especially by those in government who saw their primary accountability to the client, not the consumer. To me, this standard positions evaluation as a vehicle for improving an organization's ability to meet the needs of its consumers. While the standard appears to address only intended, direct consumers, the language of "full range" opens a window to considering indirect consumers as well. It also flags issues of equity and access to determine whether the *full* range is served. While the standard does not explicitly mention cultural dimensions, such considerations are congruent with this standard. I would remove the term "targeted," as it communicates inequality and creates a visual image that contradicts the intent of the standard.

Overview. The overview immediately broadens the focus beyond intended direct consumers ("targeted participants") to include community and society. It emphasizes the social change consequences of evaluation (see U7 Evaluation Impact) and it explicitly goes beyond evaluation as a tool of management or administration to raise questions of public good. This is consistent with AEA's Guiding Principles. The overview is terse and could be expanded to discuss direct and indirect consumers, the differing perspectives of managers, providers and consumers, and the conceptualization of public good in more detail. (Overviews for other standards are two or three times the length of this one.) ◙ This standard has several places where it could be revised to include emphasis on cultural competence, and community, participant and stakeholder involvement in the evaluation process.

Guidelines. (A) If the standards continue to apply only to educational and training programs, this language is appropriate. But program excellence should not be taken to imply a commitment to the status quo. Program excellence must be understood to extend beyond the program that currently exists to envision the best possible solution to the educational or social problem in question. Scriven's notion of a *critical competitor* is key here. ◙ Guideline A, add at the end ", and that are appropriate for the participants and community served." ◙ (B) seems to me to miss the mark. It's not a matter of informing stakeholders so much as it is working with stakeholders to discuss the purpose of the evaluation, place it in social context, and assure that the nature of the need, problem, or challenge to be addressed by the program is fully understood from multiple perspectives. ◙ Guideline B, replace with "Engage stakeholders in determining the appropriate purposes of the evaluation for the community and participants served." ◙ (C) incorrectly pairs impact on participants with organizational goals. These perspectives are *not* synonymous and interchangeable, a distinction that lies at the heart of this standard. When organizational goals and consumer needs are congruent, there is no conflict, but when the organization's goals work apart from or against consumer needs, this standard directs the evaluator to focus on the consumers. This message is lost in the translation of

Guideline C. The reference to F3 Cost Effectiveness here seems equally misplaced. (D) Though the focus on intended and unintended consumers as well as intended and unintended outcomes is relevant to this standard, the grammatical structure of D makes it hard to discern which is intended. (E) is true to the intent of the standard. It should be the lead Guideline. (F) is odd for this standard. It appears to reflect the commitment of at least some of the authors to keep the focus on improving the program in question rather than addressing consumer need. Minimally, it should reiterate, "improvement in addressing consumer need." Maximally, it should be rewritten to make the focus clearer. ◙ Revise Guideline F, "Provide interim evaluation findings to the client and other stakeholders citing strengths and deficiencies and suggestions for improvement. Solicit their input for appropriate changes or inclusions." ◙ (G) contains much that is relevant but it is written from a "top-down" position similar to B. It's not a one-way street of the evaluator informing stakeholders. It should be a two-way dialog about how the evaluation is serving the public interest and promoting the best interests of consumers. "Best interest of organization's constituents" could be read as the best interests of management and administration, and this is not the intent of this standard. ◙ Guideline G, replace "inform" with "confer with" (thus, suggesting engagement with stakeholders rather than simply reporting). ◙ (H) is a Guideline appropriate to F1 Practical Procedures; it does not fit here. A Guideline should be added (I) that addresses matters of equity. Evaluators should consider the cultural dimensions relevant to the participants and the public in the program context and examine the delivery and effectiveness of services in meeting the needs of these persons across dimensions of difference.

Overall, the Guidelines for this standard appear to have been written by persons not particularly supportive of the standard, who sought to limit its scope. This is a particular concern for our Committee, since this standard embodies many of the social change agendas that we support. The Guidelines should be rewritten to present this standard with the same depth and enthusiasm as the other Propriety standards.

Common Errors. (A) Certainly, this is an error, but one that is quite obvious and bland. The broader failure in relation to this standard is failing to consider the needs of the consumers and the impact on the public in monitoring the effectiveness of programs. One could monitor effectiveness from a narrowly prescribed, administrative perspective and not be in compliance with this standard. (B) addresses this concern appropriately. (C) is accurate but vague, not specifying why this is a failure of this particular standard. Minimally, perspectives of the providers, consumers, and community or public-at-large should be mentioned. U1 Stakeholder Identification and U4 Values Identification should be referenced. In using the word "premature" (D) implies that conclusions were reached before effectiveness could properly be demonstrated or sorted out. That would certainly be an error. But I also think that there needs to be explicit recognition in this standard that the best interest of consumers and the public are not always served by staying within the framework of an existing program. (E) begins to get at this, and its juxtaposition with D is good. (F) is entirely appropriate, but it might be more effective to introduce the link to A12 Meta-evaluation as a positive theme under Guidelines than as a negative Common Error. The multiple levels of consumer interests—direct and indirect—come across clearly in F (with age as a dimension of cultural difference implied in the distinction between students and adult learners). (G) does not go far enough. In the context of this standard, it is also imperative to consider state-of-the-art information on the nature of educational problems (social, organizational or instructional) and service delivery. If the Standards are to apply beyond education, then the current knowledge would involve the conceptualization of social problems (evidenced at micro, mezzo and macro levels) and state-of-the-art knowledge of best practices in human service delivery. **(H) is very problematic.** It labels advocacy for rights of participants or community as bias on the part of the evaluator without addressing the parallel concern of evaluators biased toward the perspectives of management and administration. **As written, this could be used as justification for judging many standpoint methodologies as bad evaluation.** We must resist the politically conservative undertone of many of the Standards and insist that they be written in an even-handed way that is respectful of multiple epistemological and methodological evaluation perspectives. (H) could be rewritten to call attention to the error of failing to consider a full range of perspectives in defining consumer need, citing P5 Complete and Fair Assessment as well as U4 Values Identification. (I) is a Common Error appropriate to F1 Practical Procedures; it does not fit here. The real complexities of this standard have not begun to be addressed. For example, the errors of : (J) Treating consumers as a homogeneous group, failing to capture differences among them with respect to need, service accessibility and service effectiveness; and (K) Assuming that program and public interests are necessarily antithetical. ◙ Common Error H, revise to read, "Failing to recognize bias on the part of the evaluator, which can lead to cultural bias and the misrepresentation of program findings and problems in relating to participants and the community at large." Also it should be emphasized that advocating about the rights of participants are appropriate when accurate.

◙**The political tensions surrounding this standard are clearly visible in the list of Common Errors, which are superficial and contradictory, and some of which undermine the very integrity of the Standard itself. This is an extremely important standard, and the Joint Committee should strengthen the language upholding it.** The current presentation is weakly supported and extremely superficial. It fails to capture the full complexity of issues surrounding the role of evaluation in supporting the public good.

Illustrative Case 1—Description. The evaluand in this case is presumably an in-service training program, the subject of which is a ninth grade math curriculum. The boundaries of this become fuzzy as the description unfolds and the evaluator is faulted for not evaluating the math curriculum itself and its impact on student standardized test scores. Rather than being an appropriate illustration of P1, it is a likely violation of A1 Program Documentation.

The context of this case is described as an urban school district, with no further information given on cultural variables. Here as elsewhere in the Case Illustrations, omitting cultural content sends the unfortunate message that cultural context is irrelevant to proper use of the Standards.

Illustrative Case 1—Analysis. The analysis correctly points out the disconnection between the scope of the evaluation and the decision made by the Division of Instruction. Though U3 Information Scope and Selection is not cited, it should be. The analysis points to several additional areas of difficulty in the case illustration: the evaluator was inexperienced; the Division of Instruction did not review proposals to determine their adequacy of scope; the evaluation was under-resourced in terms of both time and budget; the mathematics director was not conscientious in responding to the evaluator's request for input to and feedback on her design.

The analysis continues to confuse the impact of the training program with the impact of the math curriculum itself. The evaluation reported preliminary evidence that the training program led to classroom implementation of the teaching strategies. The scope of the evaluation would not permit one to draw causal conclusions concerning the subsequent low performance of students on standardized math tests. The low scores could reflect inconsistent implementation of an appropriate curriculum or solid implementation of a curriculum that happened not to be well aligned with the standardized test. The analyst suggests additional evaluation strategies that would have provided greater depth. All exceed the resources allocated to the study and many focus on the curriculum itself rather than the in-service training. Nowhere does it indicate that the evaluator was asked to examine the alignment of the math curriculum with the content of standardized tests. Student outcomes were operationalized as test performance, a point apparently supported by the analyst who does not address the evaluator's concern that an outcome evaluation was premature, pending documentation of implementation. At best, test performance is a narrow representation of consumer (student) need.

While fraught with difficulties, this case is a weak illustration of Service Orientation and should be replaced with *two* more relevant cases. To maintain a balanced perspective, I would recommend that one of the illustrations show consumers and the public best served through program improvement and continuation and the other illustrate a case in which program survival was not likely to meet consumer need and the evaluation led to a rethinking of the problem and how best to approach it.

This is an extremely important Standard from the perspective of the Diversity Committee, but it is weakly written and thinly illustrated, diluting its impact. This standard should be strengthened by deepening the Overview discussion, rewriting Guidelines and Common Errors in ways that illustrate the complexities of this standard, and adding two contrasting Case Descriptions and Analyses.

◙ Add a Case to showing how an evaluator's inaccurate interpretation of the community and participants being served was based primarily on a lack of understanding of the culture.

Supporting Documentation.

Chen, H. (2002). Designing and conducting participatory outcome evaluation of Community-based organizations' HIV prevention programs. *AIDS Education and Prevention (14)*, 18-26.

Henry, G., & Julnes, G. (1998). Values and realist evaluation. In G. T. Henry, G. Julnes, & M. M. Mark (Eds.) *Realist Evaluation: An emerging theory in support of practice*, *New Directions for Evaluation,* No. 78 (pp. 53-71). San Francisco: Jossey-Bass.

Hopping, D. (2001). Building collective capacity: New challenges for management-focused evaluation. *Children and Youth Services Review, 23*, 781-804.

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

Return to Table of Contents

P2 Formal Agreements.

Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or formally to renegotiate it.

Standard. This standard is itself culturally bound and must be understood and interpreted within cultural context. While it reflects the *modus operandi* of many bureaucratic institutions and organizations within this country, it is not necessarily congruent with all evaluation contexts to have a written agreement, nor do all evaluation models include such. The intent to establish parameters on which there is agreement among the key players is sound, although such parameters may lack the specificity envisioned by the author of the standard in the case emergent models of evaluation.

Overview. The goal of mutual understanding is clear, and the discussion of differences between external and internal evaluation and the options of a formal contract versus a memo of understanding is good. Cultural differences in modes of negotiation and/or documentation of agreement should also be acknowledged in the Overview. The Overview asserts that formal agreements should be negotiated in “an atmosphere of mutual respect and confidence.” It does not address the appropriate course of action when this condition is not present, nor does it introduce issues of prejudice and power differential that may surround and infuse the contracting process. The overview presumes a preordinate evaluation design (in which “the total evaluation plan” is known in advance) and a written final report (to which the contract can be appended, p. 88).

◙ First we need a phrase emphasizing the need to for the evaluator to take cultural differences into account when making agreements. Second, before creating the agreement the evaluator should clarify the intent and purpose for establishing a formal agreement with the client group as well as stakeholders. All stakeholders are not familiar with written agreements. In many cultures, a man’s word is his bond and the request for “paper” is considered insulting. Also in reality a piece of paper does not guarantee adherence to the conditions of the evaluation.

Guidelines. (A) alludes to making “appropriate adjustments” for emergent designs, but in fact this is a preordinate checklist that would not be particularly helpful under an emergent model. Separate illustration of formal agreements under an emergent design would be helpful. The list of areas of agreement also does not reflect considerations of cultural competence. (B) and (C) appear appropriate if there is a formal agreement. (D) should make reference to tribal laws along with the other legal examples listed. (E) “Clarity and soundness” with respect to cultural congruence may not be visible to someone outside the cultural context, so the notion of an “outside party” should be approached cautiously. The guideline seems to imply legal review, but there are other criteria by which an evaluation agreement may also be reviewed; more than one reviewer may be desirable. The reviewer could be a cultural guide who is outside the formal evaluation process, for example. (F) The explicit advice to collaborate with administrators or management (to the exclusion of other stakeholders) seems too narrow, insuring a top-down evaluation that privileges and administrative perspective. Depending on the context, the nature of the authority structure may vary, and policy development should be undertaken within culturally appropriate systems of governance.

Common Errors. (A) If one is developing a formal agreement of the kind envisioned by this standard, it would extend beyond matters of design, I agree. (B) opens a window for including matters supporting cultural competence as “important contractual elements.” (C) is extremely important and congruent with primary inclusion of consumers in the evaluation process (Madison, 1992). I like the cross-references included to U1 Stakeholder Identification, U2, Evaluator Credibility, and F2 Political Viability. (D) is excellent, raising issues of voluntary participation (P3 Rights of Human Subjects) and assumptions of power and authority that may underlie formal agreements, calling for these assumptions to be brought into the open and included in contractual negotiations. (E) is also appropriate but should be extended to include any collaborative arrangement that was previously agreed upon. Though the contractual agreement is between client and evaluator, if that contract includes, for example, the formation of an advisory group with whom the evaluator is expected to consult, then acting unilaterally in the absence of advisory group collaboration would be an error. (F) Procedures for handling cost overruns or design modifications should be spelled out in the original contract. Changes in the scope of the study would appear to require a more fundamental renegotiation. (G) I agree with the spirit of not being hamstrung by a contract, but again, the context dictates what constitutes “common sense” or an “undue delay.” These parameters may be culturally defined and not necessarily obvious to an external evaluator. Under (H), I would cite F3 Cost Effectiveness in support of the comments already made.

Illustrative Case 1—Description. The evaluand in this case is not a program, but rather “accountability procedures used by school districts,” and the client is a state teacher education association. Since the evaluand itself involves evaluation procedures, this entire case is meta-evaluation (A12). The case describes violations of several standards, prime among which appear to be F2 Political Viability, P6 Disclosure of Findings, and A11 Impartial Reporting, but the omission of final editorial authority and procedures for report dissemination from the formal agreement is a flaw that ties it to P2. This case presents no cultural information in terms of either the state teacher education association, the community, the district, or the personal characteristics of the persons involved in the case.

Illustrative Case 1—Analysis. The analysis links the violations to P6 Disclosure of Findings, but focuses on the absence of contractual agreement to protect the evaluators from the client’s unethical behavior. Outside review by legal experts is a good addition in this scenario, since the clients have already proven that they are unscrupulous. Given the behavior described in the case, it seems naïve on the part of the analyst to conclude that had a more detailed contract been in place, “No, doubt, the education association officials would have then concluded that the integrity of the report would be damaged by their requested modifications.” If they were at all concerned with maintaining the integrity of the report, they would not have removed the portions that reflected ill on them, contract or no contract. What the evaluators gain, as the analyst points out in the closing sentence, is the basis for litigation, should they choose to pursue that course of action.

Illustrative Case 2—Description. The evaluand in this case is the curriculum component of a management training program of a multistate company. The client is the company’s training director; the evaluator an external consultant who had a positive history of work for this director. In this case, there was an initial written agreement between the evaluator and the client, but subsequent modifications in the design were agreed upon verbally, with no paper trail and no explicit discussion of cost implications of the modifications. When the training director left her position, the new director declined to honor requests for payment that were not specified in the original contract, and the former director denied having agreed to any additional charges. No cultural information is provided.

Illustrative Case 2—Analysis. The analysis points to the error of not having updated the written agreement to reflect the revised design, spelling out details such as cost implications of the added component. Interestingly, the analyst attributes the error to the fact that "the former training director and the evaluator trusted one another because they had had successful dealings in the past." If indeed both parties were trustworthy, then this was a legitimate misunderstanding. The evaluator thought that the added costs had been approved; the director had apparently not realized that adding a survey would increase the cost. If the addendum had been put in writing, the misunderstanding would have become visible and been negotiated and settled. The analyst does not raise the possibility that the director deliberately lied to save face, but this too could be prevented by having the agreement in writing. The analyst raises the interesting point that the evaluator would have been better protected had more people been involved than just the evaluator and the director. The notion of building in witnesses to any agreements—apart from any substantive input they may provide—is an interesting one, although I strongly suspect that one may run into obstacles in terms of organizational culture or authority structure in certain contexts. A7 Systematic Information Control should be cited in support of the need to monitor agreements.

Supporting Documentation.

Lynch, K., Geller, S., Hunt, D., Galano, J., & Dubas, J. (1998). Successful program development using implementation evaluation. *Journal of Preventative & Intervention in the Community, 17*, 51-64.

Madison, A. (1992). Primary inclusion of culturally diverse minority program participants in the evaluation process. In A. Madison (Ed.), *Minority issues in program evaluation, New Directions in Evaluation,* No. 53 (pp. 35-43). San Francisco: Jossey-Bass.

Return to Table of Contents

P3 Rights of Human Subjects.

Evaluations should be designed and conducted to respect and protect the rights and welfare of human subjects.

Standard. The standard is a very important one, but the phrase "human subjects" comes packaged with a lot of assumptions regarding epistemology, position of the researcher and researched, etc. I'd flag it to be reworded as "participants in the evaluation process" or some similarly generic but respectful phrase.

Overview. Again, I would remove the phrase "human subjects" from the first sentence, though it is appropriate to address power differentials among participants head on by drawing particular attention to the rights of persons who are recipients of goods or services from the program being evaluated. Their role may render them more vulnerable to coercion or exploitation than that of other stakeholders who may also be participants in the evaluation. The discussion of rights and protections in the first paragraph of the overview appear to focus on program consumers or recipients as participants. Civil rights should be explicitly mentioned in the discussion of legal protections, as should accommodation mandated by the Americans with Disabilities Act.

The overview in its current version focuses on legal rights and protection, but the standard itself gives equal emphasis to **respect**. This is extremely important and often culturally defined. A paragraph should be added to the overview to examine the importance of communicating respect to participants in the evaluation, both individually and collectively. Also, as written, the focus appears to be on individual participants rather than communities or groups. Respect at the tribal, community, or group level should also be considered, as should respect along lines of cultural demarcation (e.g., Deaf Culture, religious affiliation, etc.). F2 Political Viability might well be cross-referenced here.

◙ The overview talks around cultural factors without specifically mentioning them. Too much emphasis is placed on legal consequences.

Guidelines. A record number of Guidelines for this standard! (A) is appropriately placed as the opening guideline. There's still something that bugs me in the language, "make every effort to understand"—some sort of implied condescension or power differential that puts the evaluator above the participants (sort of a throw-back to the "human subjects" concept). Does it strike anyone else this way? I think we need a guideline that prompts the evaluator to examine his/her *own* values and cultural assumptions, noting similarities and differences between the evaluator and the participant positions. (B) is appropriate, but I wonder if examples should be given of particularly relevant laws (e.g., those that establish rights of persons with disabilities or those prohibiting discrimination on the basis of sexual orientation). Or perhaps the guideline should direct the evaluator to become familiar with the laws that offer particularly relevant rights and protections to the participants of the program being evaluated in the jurisdictions that apply. As written, B is so general that it could easily be ignored. (C) is getting closer to what I had in mind: a guideline that directs the evaluator to attend to context in deciding which ethical and legal principles are pertinent. I would also cite culture here as one way to determine "pertinence," cross-listing A2 Context Analysis. (D) should be broadened beyond "instruction" to include other services or interventions. Even if the Standards remain focused on only educational programs, their scope is broader than only instructional or training interventions. (E) is headed in the right direction; however, as discussed in P2 Formal Agreements, formal written agreements are not always culturally appropriate. The guideline should specify that clear agreement be reached and documented in a manner culturally appropriate to the evaluation context. It is also important that the procedures agreed upon be clearly communicated to the participants themselves via appropriate documents or communicative strategies. (F) is written with the assumption that the participants are children and their parents indirect consumers. This underrepresents many types of educational programs. Also, is "language minority" an appropriate term for "non-English."? Could we not say something like "Assure communications are appropriate for the written or spoken language of the participants and other intended audiences." I think it would be good to mention not only linguistic translation but ASL and Braille (or its modern-day equivalent? I'm not clear about what's correct here). (G) seems pretty archaic to me. I take it to mean that the evaluators should notify participants that they will be expected to participate. The language may be appropriate for mandated participation, but this guideline itself does not communicate respect nor the voluntary nature of participation in many contexts. (H) could delete "subjects" and elaborate on clarity and culturally appropriate communicative strategies. For example, when participants are young children, what are the proper strategies for communicating the nature of the evaluation to them and to their parents? Interesting that this entire standard never uses the phrase “informed consent,” though it is hinted at in (Q). (I) Delete “subjects” and substitute participants, clients, or consumers. Literacy issues should be addressed to assure that all parties understand what they are signing. (J) I agree with the importance of gaining explicit permission rather than assuming it in the absence of objection, but here again, literacy considerations, clarity of communication, and cultural appropriateness of the communicative strategies should be built into the guideline. (K) and (L) are routine protections that are certainly appropriate, although L is sometimes difficult to assure at the group level even though individual identities are protected. (M) is never foolproof, but it certainly is something that evaluators should guard against. The more interesting question may be what, if any, are the rights of participants when data *are* misused? This engages longstanding conversations about the limitations of evaluator responsibility for data use/misuse. Not anything that can be neatly wrapped up in a guideline, but a relevant conversation nonetheless. **(N)** strikes me as odd—like a fragment taken out of context. **It doesn’t really make sense to me.** How can someone carry out the program but not participate in its implementation? The two seem synonymous to me. And even if a program provider were not included in the early planning of the program, I’m not clear in what sense they would have a “right” to withhold information on program effectiveness. It seems to me that one would have to know the contractual agreement for the evaluation (P2 Formal Agreements) to sort out the meaning of withholding information. Am I missing something here? (O) is appropriate but too cryptic. The evaluator should be directed to seek out institutional oversight procedures appropriate to the evaluation context, including but not limited to an Institutional Review Board (IRB) or human subjects committee (Are they still called this in some circles?) and to submit proposals for necessary review. Again, this procedure assumes a preordinate design, so some thought should be given to what oversight mechanisms provide comparable protection under emergent designs. (P) really only scratches the surface of individual identification. Demographic categories on a survey must be chosen carefully, for example, to avoid unintentionally revealing identities of individuals when running descriptive statistics with a small sample. An added guideline on protections appropriate to tracking respondents/non-respondents or participants/dropouts would be helpful. (Q) is phrased in a way that puts considerable distance between the evaluator and the program (the program providers are not positioned as collaborators in the evaluation but as persons being told the evaluation purpose as determined by others) which is realistic in some, but certainly not all contexts. Minimally, it would seem appropriate to acknowledge the possible (probable?) existence of multiple purposes.

Common Errors. (A) would be strengthened by giving a succinct definition of each term, since presumably, persons making this error misunderstand the terms to start with. (B) I strongly agree with this one! I would even broaden it beyond those two specific risks to speak of the general error of evaluators minimizing risk and offering false assurances of protection beyond what can be guaranteed. Promising confidentiality or anonymity when it cannot be guaranteed could then be cited as examples. (C) also speaks of false assurances. I would argue that it’s not only a function of unintended legal uses of information, as the guideline is written but also of misuses that one cannot fully protect against. The error is in overstating the boundaries of protection that evaluators can control or promise. (D) appears to refer to intended uses (as opposed to the unintended uses of C) but it's not clear to me who the implied audience is here. If it refers to clear communication with the participants, and if participants are the judges of what is "clear" then I agree. but Assuming that all purposes are overt and known. (E) seems to mean program providers as "participants" here. This error should cross-reference P2 Formal Agreements, which spoke to this in its guidelines, and also P6 Disclosure of Findings. (F) a no brainer, but it seems tautological (It violates the rights of human subjects to choose methods that violate the rights of human subjects.) Besides the obvious language shift away from "human subjects," it seems to me that the overarching error here is failing to recognize context-relevant risks and potential violations and to guard against them in the selection of methods and procedures. (G) introduces age as a dimension requiring particular safeguards. This is relevant but insufficient. The error should be extended to include persons rendered vulnerable for other reasons (e.g., health status, disability). Current IRB procedures usually specify vulnerable populations for which extraordinary protections should be enacted. (H) is, I believe, referring to the flip side of the tracking issues that I mentioned above in my comments on guideline P. I would prefer to see more detail on both sides of the argument to guide evaluators' understanding of circumstances that render at least limited identification necessary or desirable. (I) is extremely important, but I'd like to think of more appropriate and inclusive wording than "language minority."

Illustrative Case 1—Description. This case is framed as a comparative study of open versus self-contained elementary classrooms. No information is given about the cultural context of this district, though it does state that the evaluator is from outside the district. Students were to be stratified based upon achievement, personality and socioeconomic status and random assigned to the two conditions. No justification is given for the choice of these variables over others, but the design was presumably approved with the proviso that the Superintendent or her designee must sign off on any measures administered to students. Presumably under the press of time to complete the group assignments and begin the study, the evaluator administered the three measures needed for stratification—an achievement test, a personality scale, and a socioeconomic status questionnaire—without seeking Superintendent approval. No mention is made of seeking parental approval or of what assurances of protection were offered. IRB review is also not mentioned. The evaluator subsequently released the personality data of an individual child to a site principal upon the principal's request and for purposes unrelated to the study ("to better understand why the student was frequently in trouble"). That same student was used as an illustrative case in the final report, with sufficient descriptive information provided that the student's identity was revealed. No explanation is offered for the evaluator's choice here; the parents of the identified child sued the district.

Illustrative Case 1—Analysis. Given the multiple possibilities of this case, the analysis seems a little thin. The entire first half of the analysis focuses on the time frame, sympathizing with the evaluator's desire to stay on target, but noting that the timeline should have been extended to permit the required instrument review. There is still no mention of building time into the plan for institutional review and no indication that this was done. The analyst comments rather mildly that the evaluator "could have discussed with the superintendent the impact of eliminating the districts' review as a way of completing the June data collection on time" as if this were an acceptable trade-off. This seems to undermine the analyst's opening assertion that "In this case, all data collection instruments should have been reviewed and approved prior to their use." The analyst's stance on confidentiality is less ambiguous: the evaluator should not have released individual student data to the principal, despite the evaluator's eagerness to be helpful and enhance relations with district staff (mentioned in the case description). The identification of an individual participant in the final report is cited as an error, but the criticism seems mild compared to the offense. I am wondering about the context in which students and their families are disrespected in this way and whether this is an example of incompetence and bad judgment on the part of a single evaluator or a reflection of the organizational culture of the district and a recurring pattern of failure to attend to students' rights. I also wonder where the internal review procedures were that should have seen the report and flagged the violation of confidentiality before it was circulated (A7 Systematic Information Control). Finally, there are methodological questions that go unexamined. On what basis were these stratification variables selected, and why would one administer a questionnaire to elementary school students to gather data on socioeconomic status? (A5 Valid Information)

Illustrative Case 2—Description. This case is set in a dental school, and the evaluand is the dental chair. Chairs of different designs and from different manufacturers are being compared to discern which chair design permitted the dentist (student or teacher) to work best from behind the chair, a position of proven advantage in reducing back strain. Evaluators observed faculty and students working with simulated heads as well as real patients. This is where the story gets fuzzy. Apparently the evaluators had permission to observe a three-hour clinic session to note the positioning of the practitioner in relation to the chair. For reasons unrelated to the ongoing evaluation of chairs, the faculty decided to hire a clinician to grade the students on patient management and operating skills during this same three-hour time block. The graded case presentations require the student to introduce his/her patient and the patient's medical/dental history, presenting problem, and procedure. Because the evaluator was present to watch the chairs, he also heard and witnessed the case presentation. Patients subsequently complained that they had not be informed that their cases would be presented to the clinician for purposes of grading the student, and they also complained that the evaluator was present for the case presentation. No cultural information is given on the community context of the school or the demographics of its patient population, but there may be economic factors associated with who chooses to go to a dental clinic at a teaching school versus a private dentist.

Illustrative Case 2—Analysis. This is an interesting case, because the school itself erred in combining the student performance evaluation with the evaluation of the dental chairs, as the analysis points out. Since the evaluators were working in a medical context, however, they erred in not informing themselves of the rights and protections of patients as participants as participants in this context, i.e., the Privacy Act. They could have addressed this by leaving the room while the student presented the case. (The case description specifies that this case material was presented at the beginning of the three-hour session.) Presumably, the student was not yet performing in relation to the chair, so missing this portion of the session would not have compromised the chair evaluators' observations. The analysis raises another concern, however, that the presence of the grader compromised the validity of the observational data since the student under observation for grading purposes might not have been behaving in a "typical" way in relation to the chair—reactivity of the observation process introduced by the presence and purpose of the second observer. The latter, while certainly raising a validity issue (A5) is not a violation of human subjects' protections. Still unclear is what protections were instituted for these dental clinic patients in the first place. They appear to have been well versed in their rights under the Privacy Act, based upon the complaints filed, but had they given permission for the chair evaluator to be present at all? Had they given permission for their case to be presented to a clinician, a procedure that might be considered necessary in a teaching clinic? The risk here seems to have involved embarrassment on the part of the patients, who did not want to be discussed in the manner and company in which it was done. Broader issues of doctor–patient relationship and the attitudes of medical personnel toward patients of teaching clinics are not addressed, but these comprise the cultural context of this evaluation with which the evaluator should be familiar. Finally, the description indicates that the clinic administrator became angry with both evaluators and stopped both the student evaluations and the study of dental chairs. This raises interesting P2 Formal Agreements issues, since the scenario indicated that the chair evaluation was funded with money set aside from an alumni donor. Did the clinic administrator violate the original agreement and perhaps compromise the donor gift by stopping the study in that manner?

Supporting Documentation.

Gondolf, E. (2000). Human subject issues in batterer program evaluation. *Journal for Aggression, Maltreatment & Trauma, 4*, 273-297.

Martin, J., & Meesan, W. (2003). Applying ethical standards to research and evaluations involving lesbian, gay, bisexual, and transgender populations. *Journal of Gay and Lesbian Social Services, 15*, 181-201.

Shaw, I. (2003). Ethics in qualitative research and evaluation. *Journal of Social Work (3)* 1, 9-29.

Return to Table of Contents

P4 Human Interactions.

Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation, so that participants are not threatened or harmed.

Standard. To me, this single standard encapsulates many of our concerns regarding cultural competence—respecting human dignity and worth—but it extends considerably beyond avoiding threat or harm in the sense of P3 protections. Issues of power and privilege are at the heart of many interactions concerning evaluation, and evaluators must recognize the social justice dimensions of their work. This standard is written at a micro level (personal interactions) but it also applies at a macro (community or society) level.

◙ Reword P4. Evaluators should respect human dignity because it is morally and ethically correct, not just because it will prevent harm or intimidation. change "...

associated with an evaluation. This will encourage/facilitate participant cooperation and understanding."

Overview. The first sentence of the overview broadens the focus considerably, extending beyond persons who *participate in* an evaluation to those who are *affected by* an evaluation—a broader scope of influence. The overview correctly acknowledges that evaluations necessarily reflect positively or negatively on individuals or groups and their work, but it doesn't exactly integrate this with the idea of guarding against potentially threatening or harmful effects. The emphasis on interpersonal communication skills is appropriate, but again, only a part of the picture. Framing the standard in terms of "hurt feelings" seems to me to trivialize the importance of striving for understanding and communicating respect across dimensions of cultural difference. Culture is never mentioned in the overview. The intent to respect "a person's essential dignity" as a moral imperative is well taken, but again should be expanded beyond the feelings of individuals to examine the respect (or disrespect) shown to groups and subgroups who are culturally identified.

◙ Revise the last sentence of the first paragraph of the Overview to read, "Enhancement of interpersonal skills, communication ability, and cultural competence is vital to adherence to this standard."

◙ To me, this entire standard feels underdeveloped.

Guidelines. (A) brings culture into the conversation, along with values and language differences. Yet there is something unsettling and dated to me about the wording, "make every effort. To me, it communicates that this is something extra that one really should try to do, rather than a fully integrated piece of the evaluation that one cannot do without. A2 Context Analysis should be added to the list of cross-referenced standards. (B) is one that I can see both ways. On the one hand, learning about participant concerns certainly *is* time-consuming so it is appropriate to note this and to advocate for the time being budgeted for this activity. On the other hand, the phrase "Take the time" somehow hints at an optional activity that may also be passed over. Maybe I am being too picky here; I just wish the language were more assertive. How does it strike the rest of you? (C) is also two-sided. On the one hand, communicating "through established channels" can communicate cultural respect and is entirely appropriate. On the other hand, the "established channels" *themselves* may be exclusionary. If the intent of the item is to refer to the channels established in the evaluation agreement, P2 Formal Agreements should be cited. As worded, (D) focuses on the service delivery system, so P1 Service Orientation should be cited. But in addition to familiarity with the organizational context, familiarity with the community and cultural contexts should be added. Cross-listing of A2 Context Analysis is appropriate; that will be another key standard for us.

None of the Guidelines really tackles issues of respect for human worth and dignity and countervailing issues of prejudice, discrimination, and disrespect. This standard has the potential to be developed in much more powerful ways. In considering the positive and negative influence of evaluation on people, U7 Evaluation Impact should also be cited and revisited. P4 potentially picks up conversations on process-based influence that are not visible in the current construction of U7.

Common Errors. (A) is an error that addresses power dynamics. To me, the language of authority and subordinates suggests an organizational context; this might be expanded to community contexts. (B) appears to advocate non-discrimination, but given recent politics I also read it as anti-affirmative action. Just as there are times when one needs to over-sample a population to support valid conclusions, it seems to me that it might well be desirable to assign greater importance to some persons—e.g., those historically marginalized—in the spirit of balancing the evaluation (U3 Information Scope and Selection). If one is identifying illustrative categories, disability and sexual orientation should be added. (C) is clearly a violation. It seems to me that both P3 Rights of Human Subjects and A7 Systematic Information Control should be cited as well. (D) While *ad personem* attacks should certainly be avoided, it is also the case that negative findings will likely reflect poorly on persons responsible for delivering the program. A Guideline addressing respectful reporting of negative findings would be helpful. (E) is extremely relevant. The variables listed, along with others we might add such as oral expression, are largely culturally defined. Standards of professionalism and confidentiality for evaluators are extremely important and not particularly well developed in our profession, compared to say, clinical professionals. (F) This would clearly violate U3 Information Scope and Selection as well if information were collected but not needed to evaluate the program. As in the overview's framing of violations as "hurt feelings" here it seems more serious than "embarrassment"—communicating disrespect. I agree with the Error but would change the language. (G) Same here. The basic content is clearly a violation, but it goes beyond avoiding personal embarrassment. In this case, I also wonder whom they intend to refer to as "program participants." If this is taken to mean both program providers and consumers, then I concur, but if "participants" directs attention to consumers only, I strongly object.

Illustrative Case 1—Description. In this scenario, the evaluand is a history course for secondary school students, and the evaluators are academics from a university curriculum department. In one of the few references to cultural context, the students receiving the history course are described as "poverty-level." We are told that the evaluators operationally defined "poverty-level" but not told what their definition was, making it impossible to examine the values underlying that definition. No other cultural context variables are identified pertaining to these young people, so a single label defines them. Neither the cultural context of the "poverty-level" criterion nor the cultural content of the history course is discussed. The description does not explain how economic status was seen as relevant to the effectiveness of a history course, so the logic of the program and its evaluation is unclear. All 11th grade students meeting this definition (half of whom were randomly assigned to receive the history course) were brought together for pre- and post-testing. Data collection also included classroom observations and interviews with students. Both students and teachers felt disrespected by the unannounced visits (to both classrooms and homes) and by singling out students in a way that made their economic disadvantage publicly visible. It is not clear how this flawed design got past the university's Institutional Review Board and the School District. Not surprisingly, the relationship between the two was weakened by this negative experience.

◙ May consider replacing one of the Illustrative Cases in this standard for one that focuses on cultural competence.

Illustrative Case 1—Analysis. We should note with interest that this is the first analysis to attend explicitly to cultural diversity. In taking a strengths perspective with the Joint Committee, this could be cited as a step in the right direction. Interestingly, the violations in the scenario are so egregious and fundamental that to me it fits better under P3 Rights of Human Subjects than P4. The first paragraph of the analysis addresses procedural violations of basic rights and alternate steps that could be taken. It does not address the rationale for targeting "poverty-level" students to begin with, nor does it mention the absence of other cultural information of potential relevance to understanding the effectiveness of a history course. The second paragraph discusses the value of involving multiple stakeholders, but it does not address the arrogance of the evaluators' behavior nor how they would need to shift perspectives to engage meaningfully with stakeholders across dimensions of power and privilege. Simply bringing together the people identified is insufficient if the evaluators hold attitudes that devalue key players—secondary school teachers, children (and families) living in poverty. It would appear from the information given that these academics lacked the cultural competence needed to carry out a valid evaluation in this context, both in terms of the organizational culture of secondary schools and the culture of poverty. The analysis does not take this position, however, perhaps from a desire to avoid "embarrassing" the evaluators?

Illustrative Case 2—Description. The context here is the human resources department of a corporation, and the evaluand is staff development programs for secretaries. This is an internal evaluation, but no details are given concerning the person(s) identified as evaluator. Neither gender nor economic status is mentioned in the scenario, but both are plausibly relevant to this cultural context (the majority of secretaries being women who are paid considerably less than corporate executives, including managers of human resources. (Interestingly, the gender of the evaluator is not specified in the scenario, but the analysis subsequently identifies the evaluator as female.) The needs assessment included focus group interviews with secretaries. The focus group questions were developed with input from managers in human resources, and they were piloted with "groups of secretaries in the human resource department" (must be a very large corporation to have such a large HR department). Enter a manager in employee relations who sees no need to solicit the views of secretaries and tries to stop the study. One can only imagine what the real "employee relations" are in this corporation with someone like him at the helm. He did not want to deal with someone "stirring up" the secretaries, which was an eventual result of their having been omitted from participation in the evaluation design. Even after the employee relations manager changed his mind and supported the project, damage had been done to the organizational climate and the relationship between human resources and employee relations. Note that we are given no information on the historic relationship between these two units (or these two individuals) to put this scenario in context.

Illustrative Case 2—Analysis. The analysis focuses immediately on the inclusion of the secretaries and the employee relations unit, but it does not address broader systemic issues of organizational culture, power, authority, and status within the corporate environment. As an internal evaluator, there is a presumption of at least some familiarity with/competence in corporate culture, but the evaluator's view could be skewed by some of the diversity variables not addressed (gender, age, status within the corporation). The analysis does correctly point out that failure to include key stakeholders in the evaluation planning communicated disrespect and that this disrespect can be read on both individual and organizational levels. The analyst recommends formation of a stakeholder advisory group consisting of the employee relations manager, other key managers (number unspecified) and "two or three secretaries" appointed or elected by their peers. Larger issues of power differentials among these stakeholders are not addressed, yet these might well define the dynamics and fruitfulness of the proposed discussions. To me, this analysis approaches issues of cultural context, respect, and power, albeit somewhat superficially. It recognizes at least some of the issues of concern to our Committee, names them, and makes them visible.

Supporting Documentation.

Bebout, R., Becker, D., & Drake, R. (1998). A research induction group for clients entering a mental health research project: A replication study. *Community Mental Health Journal, 34*, 289-295.

Pandiani, J., Banks, S., & Schacht, L. (1998). Personal privacy versus public accountability: A technological solution to an ethical dilemma. *Journal of Behavioral Health Services & Research, 25*, 456-463.

Return to Table of Contents

P5 Complete and Fair Assessment.

The evaluation should be complete and fair in its examination and recording of strengths and weaknesses of the program being evaluated, so that strengths can be built upon and problem areas addressed.

Standard. This is an important standard insofar as it raises issues of fairness in evaluation. Shades of House's Fair Evaluation Agreement (Chapter 8, *Evaluating with validity*, 1980) and the late John Rawls, justice as fairness (though surprisingly neither is cited). I also see it tied to Messick's unified validity theory which centered on matters of completeness and cast construct underrepresentation and construct irrelevant variance as the major validity threats. Unfortunately, the actual text of this standard maintains a pretty narrow interpretation, held over from the first edition of the standards. This standard is potentially very rich and culturally relevant. It should be further developed.

◙ There should be some recognition that program strengths and weaknesses should be examined for different groups of participants. What could be a strength of a program for one group of participants might be a weakness for another.

Overview. The balance issue raised in the first paragraph is important. Evaluators must look equally hard for strengths and weaknesses, seeking evidence that might disconfirm initial impressions (be they positive or negative) or fondly held beliefs. Some such beliefs are culturally bound. They may include assumptions about persons with disabilities or those who live in poverty or those with particular religious beliefs, sexual orientations, or representing diverse cultural heritages. They may include assumptions that privilege some groups and oppress others. By introducing the concept of fairness, this standard takes an important step toward examining equity issues in evaluation.

The second paragraph opens with a questionable assumption in my opinion, implying that it is appropriate to set out to determine weaknesses as a primary purpose of evaluation, and that this is somehow OK as long as one also identifies strengths. (And if the reader is still reluctant to include strengths, he/she is given two pragmatic reasons for doing so.) To me, the opening assumption itself violates this standard. It may be the client's avowed intent to uncover weaknesses, but the evaluator should recognize this as inappropriate and take care in the initial agreement to create an even playing field.

The third paragraph focuses briefly on the methods used and raises the meta-evaluation issue of *their* strengths and weaknesses. Although I think the connection to this particular standard could be made clearer, the message appears to be that one can't really determine evaluation fairness until one examines how it was operationalized in terms of methods and procedures nor assess completeness absent the design details.

Guidelines. (A) does well to attend to both intended and unintended strengths and weaknesses in reporting results. I find myself wishing that the Guidelines started with issues of fairness and completeness in the conceptualization and planning of evaluation rather than jumping to reporting issues, however. **(B) is extremely important**. It is one of the few explicit references to diversity in the Standards and even though the authors may not have been thinking about cultural diversity, it opens the door. Here again, however, critiquing reports is too little, too late. The reviews of thoroughness and fairness should be conducted in the planning and design stages. The relevance of "knowledgeable parties representing diverse perspectives" remains. (C) anticipates some but not all of the reasons for incomplete data. Nevertheless, time and cost constraints may seriously impact multicultural validity and should be noted. The effects of such omissions should explore both issues of completeness and of fairness.

Given the central importance of issues of justice and fairness, the Guidelines for this standard seem underdeveloped and incompletely cross-referenced to other Standards. The three Guidelines offered seem to marginalize this standard by operationalizing it in terms of reporting functions.

Common Errors. (A) again moves immediately to reporting, casting issues of fairness in terms of data manipulation or deletion. These are certainly violations, but this bypasses more complex issues of fairness and completeness that must be addressed much earlier in the evaluation process. Furthering or protecting the evaluator's personal interests or biases appears in both (A) and (B) and seems more appropriately addressed by P7 Conflict of interest, which is not cited. Moreover, both of these Errors seem to underestimate the extent to which all interpretations and actions are shaped by partisan perspectives, biases, and interests (U4 Values Identification) that must be balanced. (C) again opens a window for cultural critique by alluding to the fact that strengths and weaknesses are socially constructed and calling for consideration of "alternative perspectives." As with all of the P5 Common Errors, C is framed too narrowly as a reporting concern. (D) just seems like bad method, period, violating A5 Valid Information and A10 Justified Conclusions. (E) and (F) express the same ideas as Guidelines C and A above, expressed in the negative.

The focus on reporting got me curious about the origins of this Standard, so I consulted the first edition and was reminded that this standard was originally called Balanced Reporting (C7). When the standard was reworded and in my mind broadened for the second edition, the Guidelines and Common Errors (termed Pitfalls in the first edition) were not revised and broadened beyond reporting concerns. The first illustrative case and analysis also carried over with minimal change.

Illustrative Case 1—Description. The evaluand is a two-week workshop on team teaching. The setting is a school district, but no other information is given about the providers (trainers) or consumers (teachers) of the workshop or about the "evaluation group." A committee from the school district asked evaluators to find weaknesses in the workshop materials and suggest improvements. The evaluators conducted post-workshop interviews with teachers and trainers to inventory perceived weaknesses and suggestions for improvement. Apparently the developers of the materials were separate from the trainers, and these developers found the recommended changes to be in conflict with the core characteristics of what they had designed. The report was not useful in revising the materials. No cultural context information is provided, and this case does not illustrate concerns about fairness.

Illustrative Case 1—Analysis. The analysis does not address the failure to consult with developers as a stakeholder audience, but notes the obvious omission of data on strengths to balance the data on weaknesses. The analyst then proceeds to suggest a different evaluation design in which materials are rated by section for effectiveness and usability, the process of workshop delivery is monitored and teacher skill and attitudes are assessed. While this may be a reasonable design to apply to such an evaluand, it is not responsive to the stated purpose of the evaluation nor is that purpose renegotiated to support the broader design. It seems to me that a caveat here is that expanding the design in the spirit of completeness may conflict with the stated purpose of a given study.

Illustrative Case 2—Description. This case presents an external evaluator from another state who is called in by a university administrator to evaluate a professional degree program. Other than the fact that this is a state university, no information on context is given. The evaluator was given program information prior to a site visit that covered one evening and the following day. (The scenario states that these program materials were "hastily assembled," a puzzle since it also states that use of external reviewers was commonplace in this state system.) The site visit involved a series of individual and group interviews with a range of stakeholders, and the evaluator was asked to completed a 60-item standardized rating form used by all state universities in this state. The evaluator expressed concerns about content validity of the instrument for this particular program and about the scoring procedures, which gave all items equal weight. He expressed these concerns in his site visit debriefing with the university chancellor and in his written report.

Illustrative Case 2—Analysis. The analysis points to the need to clarify the full range of purposes of the report at the outset, rather than learning about the comparative use of his review during the site visit. The analyst asserts that a clearer sense of purpose would have permitted collection and analysis of relevant data "to allow for more complete and therefore fair reporting." The equation of completeness with fairness here is noteworthy; I understand them to be two different constructs, but the standard never really tackles what I would consider to be issues of fairness. Interestingly, the analyst also refers to details not included in the description. For example, he/she notes, "the large groups of students and alumni and the setup of the room made it difficult to assess their perceptions about the adequacy of the program," but the reader is told only that students and alumni were each interviewed in two groups, nothing about group size or room setup. Although required to use a standardized rating form, the evaluator fleshed out the report with additional information to improve balance. The inclusion of the addendum in the statewide review should have been clarified in the initial negotiation (P2 Formal Agreements). The analysis does not comment on the validity concerns about the instrument nor the broader validity concerns about completeness and fairness.

Supporting Documentation.
House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.
Rawls, J. (2001). *Justice as fairness: A restatement.* Cambridge, MA: Belknap Press.

Return to Table of Contents

P6 Disclosure of Findings.

The formal parties to an evaluation should ensure that the full set of evaluation findings along with pertinent limitations are made accessible to the persons affected by the evaluation, and any others with expressed legal rights to receive the results.
Standard. This standard supports inclusion of persons outside the normal authority structure by specifying that the full findings and their limitations should be accessible to "persons affected by the evaluation." This could include economically or politically marginalized groups, direct and indirect program consumers, those from who information was collected during the evaluation, and the public. "Expressed legal rights" includes the Freedom of Information Act. This is small standard with large impact. Information is power, and by assuring access to information, this standard opens up the opportunity for cultural critique by affected parties. (Interestingly, this standard was originally titled "Public's Right to Know" (C4) in the first edition and had an even clearer advocacy stance, in my opinion.)
◙ What about the non-formal parties? All members of the affected community should have access to the findings.

Overview. In the first paragraph, reference to public safety and the abridgment of individual freedoms now takes on an entirely different meaning than intended in the 1994 edition. It will be important for the Joint Committee and possibly our own **AEA Ethics Committee** to examine the implications of Ashcroft-era legislation and judicial interpretation on access to evaluative information. The current overview makes a strong and explicit statement of support for disclosure that is direct, open, and complete (paragraph 2). To my reading, this standard takes a strong advocacy stance that is wholly congruent with the values of inclusion and supportive of multicultural validity. It connects disclosure to issues of both utility and fairness. We should keep our eye on this one to insure that its intent is not diluted in future editions.
Guidelines. (A) picks up the last point of the Overview and expands it, appropriately citing P2 Formal Agreements. Certainly, this is an important dimension; disclosure parameters and procedures should be included in evaluation contracts or understandings. I would have personally preferred to see (I) come first, because that sets the broader context for contractual negotiation around disclosure. (B) is interesting insofar as the means of reporting will differ across evaluation models. I have mixed reactions to the guideline specifying written communication, because I can envision situations in which

this might not be culturally appropriate or congruent with a particular model (e.g., Patton's Utilization-focused evaluation). On the other hand, in the spirit of full disclosure, it is hard to argue that findings could be accessible to all persons affected without any sort of paper trail. An interesting dilemma. (C) is very important to issues of diversity—e.g., the justification for inclusion of race as an explanatory variable (Davis, 1992) or any similar procedures for including or excluding cultural dimensions should be fully justified. (D) calls for balance and fairness, as does (E) though the connection to P5 Complete and Fair Assessment is not explicitly made. I would prefer the term "fair" instead of "broad," as used, because some judgments or recommendations may be narrowly focused. But I agree with the intent, which is that a broad base of information, taking multiple perspectives, support a recommendation. (F) is extremely important, citing A2 Context Analysis, which can incorporate cultural context. Because it is broadly framed, I think this guideline should also cite A12 Meta-evaluation. (G) also seems quite specific to particular types of studies, rather than a statement of general principle. I guess "where appropriate" is the key here, but it still feels pretty constraining. I would prefer to see a Guideline cautioning against thinking of disclosure only in terms of a final report (or maybe that's one for Common Errors). (H) focuses on the aspect of this standard that fosters inclusion and is therefore one that I would move earlier in the list. Culturally competent communication is not limited to "appropriate linguistic form," though certainly that is a beginning. (I) will need to be revisited, updated and expanded to address current legal and political climates affecting civil liberties and to bring in regulations regarding electronic communication, which appear nowhere in the current edition. (J) is an important override to (A); for greater impact, it should be positioned immediately after (A).

Common Errors. (A) should be expanded to include additional rationales for inclusion—e.g., social justice, equity, cultural diversity. Alternatively, a separate Error could be added to address lack of diversity among stakeholder audiences restricting the understanding of context, strengths and limitations, and other matters cited in the Guidelines. (B) refers to the evaluator failing to exert control? Not completely clear who is the intended referent here. It seems appropriate to cite A7 Systematic Information Control here too. (C) and (D) are specific types of contracting errors in the agreement between client and evaluator(s). (E) raises interesting issues of equitable disclosure within rather than across groups. If an evaluator respects the lines of authority within a particular cultural group and reports, say, to the elected leader of the group, he/she may not be reaching factions of the group that hold opposing opinions. I am picturing the tribes of the Oneida nation nearby and the presence of different political factions, all of whom would be meet the criterion of "affected by the evaluation." Achieving equity could be very tricky. This one may need to be cross-listed with F2 Political Viability. (F) As written, altering a report in this way would be a clear violation, but it may well be the case that there are subtleties of emphasis, perspective or wording that reflect these particular vantage points. It would be an error not to examine the report for such subtleties, making this Error more complex and less obvious than it may first appear. (G) "Premature" implies that the evaluator would have made the information clearly interpretable and understandable, given more time. To me, it seems as though timing and clarity are two different matters and are best addressed separately. (H) is interesting on several levels. First, like (B) it doesn't really explain who is being addressed here. An explicit role of meta-evaluation *is* to make both strengths and limitations visible. If the limitations are extensive, the evaluation may well deserve limited credibility. The notion of "making too much" of the limitations sounds as if it's OK to cite a couple weaknesses and then let it go at that; it does not encourage serious meta-evaluation (A12). If balanced reporting is the key point here, then the Error should be rewritten to reflect that. (I) is extremely important but must be updated to reflect current debates about privacy. (J) raises important matters of clients' rights, responsibilities, and needs, appropriately citing P4 Human Interactions, but the wording "be considerate" seems a little weak to me. "Respect" or "comply with"—any language that makes clear that we are speaking of rights to be upheld here, not matters of politeness or social graces. The wording should send a strong message that these are not optional considerations but mandated areas of concern. (K) creates an opening for noticing failures to consider cultural context. My only concern is that it comes last in quite a long list and may get lost.

Both the Guidelines and the Common Errors in this Standard seem to be spelled out in finer detail than for other Standards, often in ways that would not be universally applicable to all evaluations. The Joint Committee may wish to attend to level of detail in its editing of the next edition, so that a comparable level of generality or specificity is maintained across Standards.

Illustrative Case 1—Description. This is one of only a few cases with explicit cultural content. It is carried over with slight revisions from the first edition. The evaluand is a school district's desegregation strategies. Presumably, the reference is to racial desegregation, although this is never stated. The factions are ambiguously identified as "majority group" and "most prominent minority group" with no sense of diversity within groups nor any further information on the cultural context of this community (though the community becomes an important stakeholder as the scenario unfolds). The three-person evaluation team is external to the school district; no information on their racial or ethnic composition is given. The formal agreement concerning disclosure puts control in the hands of the Board with respect to timing, content, and the persons to be informed, but the language of informing persons "in due course" and at their discretion makes no commitment to full and fair disclosure. The evaluators, for reasons unstated, chose to focus the study on the students and the school system, to the exclusion of parental perspectives and those of the community, despite becoming aware of strongly divided opinions on desegregation among the citizenry. Before releasing the final report, the Board requested that content on controversial bussing decisions be edited out, and the evaluators complied. Community stakeholders did not receive the report well, challenging their exclusion from the evaluation process and the omission of controversial information from the report. They accused the evaluators of violating public disclosure laws.

Illustrative Case 1—Analysis. The analysis points to the importance of insuring that formal agreements are in compliance with federal and state laws relating to disclosure of public information. It goes on to recommend that an advisory group be formed, inclusive of school and community, to provide input into the evaluation plan and reporting procedures. The issue of racism is never named or addressed head on, but in a noteworthy understatement, the analyst speculates that the advisory group strategy would "perhaps increase the complexity of the evaluation." Intermediate reports are recommended to address specific issues. The history of bussing is used as an example, though why the

Board would agree to this is not addressed. The analyst stresses the need to explicitly seek both majority and minority input and responses to the report. The analyst upholds the public presentation of the report without alteration, "even though many of the recommendations may have been unpalatable to the majority and/or minority groups and to the board." Again, issues of racism are not named. Instead, the negative reaction to the report is framed as poor process, with the assumption that it would have been better received "if the public had been openly, honestly, and fairly informed at all stages." While this may be true as a general rule, it does not address dynamics of power and prejudice that may work against this standard at the levels of organizations, communities, or societies. Good process may make such dynamics more visible but it does not necessarily change them.

Illustrative Case 2—Description. The second case is set in the medical school of a large public university. No information is given on the cultural context of the program or the cultural composition of the faculty or student body. The evaluand is undergraduate clinical group teaching. The evaluation was internal, planned by an interdepartmental curriculum committee after consultation with "all relevant constituencies" (roles not further specified). There were two distinctly different audiences, each with its own focal questions. Curriculum management committee members and chairs of specialty curriculum areas were interested in how well the clinic groups aligned with the instructional objectives of the specialty areas. Department chairs were interested in assessing the performance of clinical faculty for purposes of promotion and tenure. Students rated both the educational content and the instructor performance at the end of each specialty unit over a three-year period, with better than 90% participation. Separate reports were prepared for each audience, and results were not released until after students' grades had been submitted. The evaluation system itself was reviewed annually and modified as necessary.

Illustrative Case 2—Analysis. The analysis highlights the ways in which this case was in compliance with standard P6: Stakeholders were identified and consulted; questions were framed around two clearly-defined information needs, specifically defined for each audience; results were circulated to relevant stakeholders without tampering; the data were released at year's end to protect students from any possible faculty recrimination. All of these procedures are appropriate, but nothing in this case or its analysis really zeros in on the potential complexities of disclosure, including opposing agendas or political positions, issues of harm in the use of information, a culture of litigation, or differing stakeholder perspectives. For example, one intended use of findings is in promotion and tenure decisions, yet there's insufficient information given to appreciate the complexities of using the data described for this purpose. I like the idea of including positive as well as negative case illustrations, but it would be more useful to show how challenges were met rather than to present a case that appears to be without such challenges.

Supporting Documentation.

Curcio, C., Mathai, C., & Roberts, J. (2003). Evaluation of a school district's secondary counseling program. *Professional School Counseling, 6*, 296-303.

Davis, J. E. (1992). Reconsidering the use of race as an explanatory variable in program evaluation. In A. Madison (Ed.), *Minority issues in program evaluation, New*

Directions for Program Evaluation, No.53, (pp. 55-67). San Francisco: Jossey-Bass.

Return to Table of Contents

P7 Conflict of Interest.

Conflict of interest should be dealt with openly and honestly, so that it does not compromise the evaluation processes and results.

Standard. This is a basic ethics issue, but it could take on overtones of cultural identification. I also think this one is a relevant reminder for those of us who are deeply committed to social justice that there are many ways in which it can be compromised, including conflict of interest at the individual level. It's good that both the evaluation process and results are noted as potentially subject to compromise; this is congruent with consideration of evaluation influence as both process-based and results-based.

Overview. The overview really zeros in on financial conflicts of interest without considering more subtle or complex issues of personal investment or benefit that may compromise an evaluation. Personal interests may be ideological as well as financial. That aspect is not well developed in the overview's first paragraph. Of the six examples provided, three relate to financial gain and three concern non-monetary benefits. This list is more balanced, though no example relates to culturally based conflicts. I strongly agree with the last sentence on p. 115 that potential conflicts of interest are essentially omnipresent in evaluation, and that the key to responsible evaluation is learning how to deal with them. Given that no evaluator can step outside of his/her own interests and values in any absolute sense, it seems that conflict of interest becomes a continuum, with the key judgment being, "When does a personal interest become 'inappropriate'?" Interestingly, the kind of ongoing self-reflection that this question implies bears similarity to the self-awareness component of most cultural competence schemata. Both of these constructs (conflict of interest and cultural competence) imply that evaluators must be self-reflective beings—an interesting assumptions and one not explicitly addressed by any Standard. On p. 116, the Overview concludes with two important points: that conflict of interest applies equally to internal and external evaluation, though born of different issues; and that conflicts of interest infuse bias throughout the entire evaluation process. This Overview raises a number of important issues, improving in its breadth as it goes on. It readily lends itself to further expansion to deal more explicitly with cultural content, an area not visible in the current version.

Guidelines. (A) While this is not an unreasonable Guideline, it is not the first step. First one must have the sensibility to reflect on roles, context, and circumstances and come to a personal understanding of the potential conflicts imbedded in a given evaluation. The word "identify" does not alert the reader to the complexity of this task. (B) may be an appropriate mechanism in many cultural contexts, but oral communication, ritual, ceremony, and other ways of securing agreement should be recognized also. I like (C), although it is framed quite generally and does not explicitly draw a connection to conflict of interest, and the cultural dimensions of "different perspectives" have yet to be drawn out. U4 Values Identification, and P5 Complete and Fair Assessment should be cross-referenced. A related point that is not tackled is that the definitions of conflict of interest themselves may be culturally bound, differing among the "various perspectives" of cultural subgroups, adding another layer of complexity to the conversation. It is significant that (C) notes the importance of "staying open" to alternatives, reminiscent of Ridley et al.'s (1994) argument about plasticity as a component of cultural sensitivity. (D) is an extremely relevant antidote to conflict of interest. I would move it even earlier in the list and cite A12 Meta-evaluation. (E) avoids certain conflicts of interest but may introduce others. I would prefer that the Guideline be reworded to direct evaluators to consider potential conflicts introduced by the funding pathway rather than unilaterally supporting one pathway over another. **(F) is extremely important, as it explicitly broadens considerations beyond monetary gain to address culturally-defined advantages.** Status, prestige, power could all be gained or lost as a result of evaluation. (G) is troublesome, introducing potential managerial bias and implying that "agency heads" are somehow free of conflicts of interest. As in (E), I would prefer that the Guideline direct attention to conflicts born of organizational position and authority rather than direct evaluators to work from only one perspective. As for (H), it seems to imply that both conditions are optional. However, as the overview suggests, conflict of interest is omnipresent, making meta-evaluation (cast broadly) an important component in all cases.

Common Errors. I agree with (A) that calling attention to conflict of interest does not assure that it will be handled properly (a distinction that parallels the difference between cultural awareness and cultural competence). One may be aware of conflict but lack the knowledge or skills to reach a responsible resolution. I agree with (B) on two levels: that one cannot eliminate all conflicts of interest in any absolute sense, and that "well-established procedures" certainly offer no guarantee against conflict of interest. If procedures were well-established in majority contexts, they may, in fact, *introduce* conflicts of interest when applied in different cultural contexts. I'm not sure that anyone would truly believe (C), but it certainly bears repeating that no one can step outside his/her own values, interests and life experience to claim immunity from potential bias. (D) is an important caveat, and I would cite A5 Valid Information because this really raises a validity issue. Exclusion is not necessarily the most appropriate way to address conflicts of interest, particularly if such exclusion compromises the validity of the study. This could apply to experiential justifications of multicultural validity, for example.

Illustrative Case 1—Description. The evaluand is an individualized reading program at the elementary level. A curriculum director and a reading specialist from the district along with "several teachers" developed the program. The curriculum director is seen as the "major architect." No cultural context is given concerning either the district or these persons who developed the reading program. The program involved three stages. The curriculum director "commissioned" a team of teachers to evaluate stage one, with the understanding that they might be called upon to continue to evaluate stages two and three if their work proved satisfactory. No financial or other procedural details of the agreement "commissioning" the teachers are given. The evaluation team involved other staff (nature and number unspecified) in developing criteria to evaluate the program, and the criteria finalized by the team focused on program strengths rather than weaknesses, presumably violating P5 Complete and Fair Assessment. When the positive findings were released, teachers complained that their objectives had not received proper consideration, that deficiencies and controversies had not been given sufficient visibility, and that one of the evaluators had also participated in program development. The superintendent concluded that F2 Evaluator Credibility had been violated and dismissed the evaluators.

Illustrative Case 1—Analysis. The analysis notes that these evaluators should have "assessed and dealt openly" with their own conflicts of interest and those of the client and other stakeholders at the outset, declining to proceed with the study if the evaluation could not be kept "reasonably free from vested interests." While no specifics are given, the analyst notes that staffing and design issues should have been handled differently. The analyst notes that inclusion of key staff persons was appropriate but should have been broadened to include the views of those who had concerns about "controversial elements" of the program. Because no cultural context information is provided, one cannot determine whether any of these controversies are culturally connected. The recommendation that external evaluators from another district be brought in to add credibility may or may not resolve the conflicts of interest, depending on the context details.

Illustrative Case 2—Description. The evaluand is a continuing medical education program (CME) offered by a professional association. The evaluators are medical school faculty from an educational research development department; they won the contract via a competitive process. An advisory group for the evaluation consisted of five members of the professional association offering the program and three "external consultants." No historical or cultural context information is given for the professional association, the physicians who designed the evaluand, the medical school, the evaluators, or members of the advisory team. Two specific courses were chosen for evaluation, but the topics and content of these courses are not described, so it is difficult to envision the changes in physician practice that were monitored via physician interviews in the comparison group design. Absent this information, it is also impossible to interpret the fact that changes reported in the interviews were not visible in patient charts. The evaluators attributed the lack of agreement to lack of validity of the charts as a criterion. The scenario does not discuss how they proposed to correct this problem in the next round of evaluation. A5 Valid Information is clearly a concern.

While this is all very interesting, it doesn't seem to me to relate to P7 until the last paragraph (pp. 118-119). When the interview data suggested that the official course objectives did not match the information needs of practicing physicians, the evaluators recommended systematic and periodic needs assessment of community physicians. The investment of the professional association in maintaining the current curriculum without change was thus revealed. It turns out that the curriculum in question was designed by "high-profile academic physicians" who exerted considerable leadership in the association despite being out of touch with the realities of community practice on a day-to-day basis. The association members on the advisory group argued that the academic freedom ("autonomy") of these faculty members would be threatened by collecting needs assessment data that might suggest curriculum change. We are given a glimpse of the personalities involved as well (personal *is* political) when it is asserted that these high profile physicians would withdraw from the association if their curriculum were challenged, weakening the lobbying efficacy of the entire association. Faced with a choice of employing only "token field-based needs assessment" or being ignored and dismissed, the evaluators resigned. No details are given on the manner in which they terminated their contract.

This is a comparatively lengthy scenario, raising a number of problems, only some of which concern Conflict of Interest. To me, it would have more impact as an Illustrative Case if it were edited to focus on P7.

Illustrative Case 2—Analysis. The analysis goes immediately to the P7 issues raised by the restrictions placed on their evaluation plan to secure the "renewal of a lucrative contract." The analyst frames the evaluators' choice as a decision not to pursue another contract, and does not address any conflicts implicit in the initial contract. He/she does suggest that the impasse could have been avoided if (1) the conflicts been recognized sooner and discussed with "association leaders," (2) early results suggesting change had been discussed via informal communication with the oversight committee and association leaders, and (3) the faculty conducting the programs had been included as stakeholders and involved in designing the evaluation. These are all plausible "textbook" recommendations, but one wonders about their fit with the organizational cultures of this CME association, the medical school (and its Institutional Review procedures), and the lobbyists, and the dynamics of cultural similarities and differences among the individual key players in the scenario. Since this information is not given, it can be raised as a curiosity but not explored. This scenario is also interesting since it explores dynamics of power among some of the most historically privileged members of US society—high profile academic physicians who are powerful lobbyists, physicians who are medical school faculty members, and physicians in community practice.

Supporting Documentation. ◙ Note the funny typo in the Windle & Neigher (1978) reference. The correct subtitle is "Advice for *trapped* evaluators" not "tripped evaluators" (which would raise an entirely different set of ethical concerns). ☺

◙ I still find Newman & Brown (1997) to be the best overall reference on ethics and evaluation.

Dial, M. (1994). The misuse of an evaluation in educational programs. *New Directions for Program Evaluation,* No. 64, (pp. 61-67). San Francisco: Jossey-Bass.

Moskiwitz, J. (1993). Why reports of outcome evaluations are often biased or uninterpretable: Examples from evaluations of drug abuse prevention programs. *Evaluation and Program Planning, 16*, 1-9.

Newman, D. L., & Brown, R. D. (1997). *Applied ethics for program evaluation.* Thousand Oaks, CA: Sage.

Ridley, C. R., Mendoza, D. W., Kanitz, B. E., Angermeier, L., & Zenk, R. (1994). Cultural sensitivity in multicultural counseling: A perceptual schema model. *Journal of Counseling Psychology, 41*, 125-136.

Vroom, P., Colombo, M., & Nahan, N. (1994). Confronting ideology and self-interest: Avoiding misuse of evaluation. *New Directions for Program Evaluation* No. 64, (pp. 49-59). San Francisco: Jossey-Bass

Return to Table of Contents

P8 Fiscal Responsibility.

The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible, so that expenditures are accounted for and appropriate.

Standard. I find nothing to disagree with in the basic wording of this standard, but I note with interest that it refers to expenditure of *resources* for evaluation, not expenditure of funds. So I would take this to mean that despite the title of the standard (unchanged from the first edition), the authors are prepared to consider both monetary and non-monetary resources, which includes people resources and may bring us around again to some P4 Human Interactions issues. I agree with this broad framing. With respect to our focus, it raises interesting issues of how culture may be viewed as a resource to be respected and "prudently and ethically expended" in evaluation. What are the appropriate (sound) accountability procedures for overseeing the expenditure of cultural collateral?

Overview. The overview unfortunately narrows the focus of the standard to financial resources and the accounting, auditing and regulatory procedures governing fiscal transactions. The overview restricts the utility of the standard by directing exclusive attention to cultural contexts in which money is exchanged. Notions of payment for services rendered should be broadened to include barter economies and responsible monitoring of *pro bono* work. As I already mentioned, I also think there are some fascinating possibilities here to explore issues of cultural collateral.

The caution against actual or alleged misuse of funds is certainly appropriate to overall ethical practice. It raises interesting issues of culturally appropriate accounting strategies, assuming that computer spreadsheets or traditional ledger books may not always be the methods of choice.

Guidelines. (A) presents a fairly traditional list of monetary cost categories associated with preordinate evaluations. It would be an interesting exercise to augment these with categories of cost likely to be incurred in delving into cultural context or educating evaluators in the history and traditions relevant to a given evaluand. These cost categories are framed in terms of "billable expenses" that one might use to structure an evaluation budget, but they should be expanded to include costs other than monetary costs and costs to persons other than the evaluators. For example, a cultural guide may or may not be paid money for services rendered depending on the context, but his/her time certainly gets used up and this should be conceptualized as one of the costs of the evaluation. Given that the costs listed are ones able to be anticipated under a preordinate model, it would be interesting to develop a parallel guideline for estimating and tracking costs under an emergent design. (B) is particularly relevant to maintain accountability to traditional external funders in this country such as government agencies or private foundations. The only caveat may be the contextually determined definition of what constitutes "clear and understandable" accounting practices. (C) has implications for monitoring both fiscal costs, in the case of paid personnel, and non-money costs, in the case of donated time or use of volunteers. Opportunity costs should also be considered, the issue being what types of records are "adequate" to permit reliable assessment of opportunity costs? (D) may be appropriate as a general rule, but going with the lowest bid may not produce culturally competent evaluation. As a common error, cultural competence should not be sacrificed in the interest of saving money. (E) While I strongly support public disclosure, this may vary across organizational, cultural, and funding contexts. I agree that such fiscal information should be available upon request. (F) concerns me not because I would want to waste money, but because of the labor-intensive nature of inclusion and other procedures that support cultural competence, I worry about placing too high a value on frugality.

Common Errors. Interesting that there are nearly twice as many Errors as Guidelines for this standard. (A) is appropriate for most traditional, preordinate evaluations. It would be helpful to have guidance about the level of clarity that is desirable up front in an emergent design. (B) Costs associated with including an appropriately broad range of relevant stakeholders should be included here. As one comes to a deeper appreciation of diversity within stakeholder perspectives, for example, additional persons may need to be included in the evaluation. This may translate into budgetary adjustments or, in the case of non-money costs, adjustments to the project timeline. (C) Evaluators must be familiar with laws and regulations governing both the reporting of financial expenditures and the documentation of donated time and goods. (D) Though this error is rather vague, I assume from the reference to P7 that it refers to favoritism in contracting for goods and services. Before fully endorsing this statement, I would want to consider the affirmative action implications. For example, if this considers it to be an error to “favor” a minority-owned evaluation firm in evaluating a program directed at a minority population, I would likely disagree. On the other hand, I am not advocating violation of P7. (E), (F), and (G) appear to refer to basic management practices. (H) is dishonest in representing the caliber of persons who will be completing the work. The Standards are right to take a firm and explicit stand against this all-too-common practice. However, in judging who is “qualified,” it is important to make sure relevant dimensions of cultural competence have been taken into account. I can envision a situation in which the senior-level staff may be less qualified for work in certain contexts than more junior staff. Qualifications or competence, especially around matters of culture, should not be equated with seniority or organizational status. (I) The definition of what constitutes “substantial” change requiring reauthorization should be included in P2 Formal Agreements. Like (H), (J) is dishonest, though it is common to allow a margin for unanticipated costs. Though money is rarely returned, expenditures should be relevant to supporting the quality of the evaluation or its dissemination, and the Standards are correct to be explicit about this.

Illustrative Case 1—Description. Another medical school example, perhaps reflecting the composition of Joint Committee members. This time, the evaluand is “an innovative curriculum for medical residents” for which funding was sought from a federal agency. Neither the topic of the curriculum nor the nature of the federal agency is mentioned, so one cannot intuit what cultural dimensions might be relevant. An evaluation plan is required, and the principal investigator asks an evaluation specialist from another department within the school to prepare that section of the proposal. Conversations occur between the evaluation specialist and the PI but for reasons that are unclear, the evaluation is under-budgeted in the original proposal (faculty are expected to commit time to the evaluation for which they are not being paid). On top of this, the proposal is not funded at the requested level, necessitating further budget cuts. The crux of the problem is that the PI did not inform the funding source of changes made in both the program and the evaluation to accommodate reduced funding. Also there appear to have been some unclarity around definition of roles and responsibilities, since it indicates that the evaluator “could not independently initiate any activities.” Necessary data were not collected, a minimalist design was substituted for the comprehensive plan that had been proposed, and the funding agency was seriously angered. They refused to consider future applications from either the PI or the evaluation specialist.

This is another case example that violates multiple standards (of common sense as well as the Joint Committee), so that the illustration of P8 gets a bit lost in the shuffle.

Illustrative Case 1—Analysis. Since neither the PI nor the evaluation specialist carefully examined the budget prior to submitting the proposal, they certainly did not demonstrate fiscal responsibility. Moreover, revisions in the evaluation plan were not developed and agreed upon in a timely manner when the project was funded below the level requested. (Seasoned evaluators should actually anticipate such action on the part of funders and design a contingency plan at the time of the original proposal development. This might be pointed out in a Guideline for this standard.) The analyst also notes that there was “poor communication”—an understatement—between the evaluator and PI. The evaluator was a woman, the PI a man, but without more details, it is impossible to sort out how gender differences may have contributed to the communication breakdown or whether other dimensions of difference were also relevant. The organizational culture should be explored here as well. Since the evaluator was from another department, there may have been historical or structural barriers to effective communication between faculty members in these units.

There is only one case illustration for this standard. When adding a second case, it would be useful to take it beyond matters of financial budgeting. If it is truly the intent of the standard to remain narrow, a separate standard should be added to address the broader issues of responsible documentation and oversight of non-money resources and opportunity costs, cross-listing F3 Cost Effectiveness and P4 Human Interactions.

Supporting Documentation.

[No resource identified]

Return to Table of Contents

Accuracy Standards

The accuracy standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the program being evaluated.

A1 Program Documentation.

The program being evaluated should be described and documented clearly and accurately, so that the program is clearly identified.
Standard. The definition of the evaluand is an extremely important component of evaluation. I'd have to cite Scriven (1991) again on this one. Understanding the meaning and logic of the program in cultural and historical terms is also an important piece of clear and accurate description.
Overview. I like the emphasis on gaining an understanding that is solid or has depth. It makes clear that this is referring to more than a superficial statement clipped from a program document or a logic model. I also like the connection to validity issues here. The need to distinguish between intention and actual implementation also underscores the role of process evaluation, which is an excellent point. Though cultural dimensions are not mentioned, it is clearly congruent with the intent of the standards that cultural meanings be included in building a solid understanding.
Guidelines. (A) probably includes consumers among "the other stakeholders" but I always prefer to see them named explicitly. Also, one may want to steer clear of compartmentalizing jargon (e.g., objectives) and ask more general questions such as what is their understanding of what this program is seeking to accomplish. Language should be chosen to reflect the communicative style and cultural context of the stakeholder. This means that it may not be possible *or desirable* to get written descriptions. The focus should be on what is expressed in the Overview as "a valid characterization of a program." Validity of the description is paramount. ◙ Guideline A: add "participants" to the list of program characteristics stakeholders should describe. ◙ (B) This is an important point that raises interesting possibilities, though the authors are still thinking in terms of written documentation or electronic records. ("Slide-tape presentations" would largely translate into PowerPoint today.) These sources could be augmented by/compared with the conversations from (A) or direct observation of things communicated by the décor and layout of the physical space. ◙ Guideline B: Note that only program-generated information is listed. Participants groups often also discuss the evaluation in community newsletters, etc. These documents should also be collected and documented.
◙ © To support the validity of their observations, these "independent observers" should be culturally competent. ◙ Guideline C: Who are the independent observers? Does this refer to "objective observers," i.e., non-participants, and if so, how do they define such a person. I agree with (D); it is important to watch the program in action. ◙ Guideline D: add "paying specific attention to contextual factors." ◙ (E) is definitely important. Ridley et al. (1994) speak of plasticity as one of the components of cultural sensitivity. Though they are speaking at the micro level of individual counseling, I think the same principle applies to the macro level—in this case, not getting "stuck" in an initial understanding of the evaluand and its context but being able to rethink description and revise it as one's understanding deepens. (F) would depend on whether one is doing a global or an analytic evaluation (Scriven, 1991) but for an analytic evaluation, this is an appropriate guideline.

(G) should be presented as an iterative process, consistent with (E) rather than as a single "check." This process is especially important to build a multiculturally valid description. (H) is awkwardly worded, but it seems to advocate inclusion of evaluand description in a written report, along with process evaluation data on actual implementation. The guideline should also recommend inclusion of cultural context in portraying the evaluand, and this should be modeled in the case illustrations.

Interestingly, these Guidelines address the importance of gaining multiple perspectives, but not how these multiple perspectives will be synthesized into a consensual understanding or representation of the evaluand. This is an important aspect that should not be overlooked.

Common Errors. (A) would certainly be an error, and it is exacerbated by the fact that such "official" descriptions typically lack the cultural context information necessary to support a full and genuine description. (B) raises an important validity issue regarding the description itself. Program consumers and other relevant stakeholders should be included in the list of those who should not be overlooked. I would add as an error, "assuming that the majority description of the evaluand represents how it is understood from all cultural perspectives." (C) minimizes contextual influences as well as failing to explain the logic of program influences. (D) is important on a couple of different levels. First, it is one of the few places in the text that emergent designs are acknowledged or addressed. Second, the notion of "forcing" a description or an understanding onto an evaluand could include cultural insensitivity on the part of the evaluator. (E) seems to create a false dichotomy, perhaps reflecting the linear thinking of preordinate designs: first you describe the program, then you evaluate it. To me, one's understanding of the evaluand evolves throughout the evaluation process and the description of the evaluand continues to be revised and updated as the evaluation proceeds. Clearly one must budget time well throughout the process, but I worry that putting this forward as an error would discourage the inclusion of multiple perspectives, which *is* time-consuming. Perhaps the error is "Viewing program documentation as an exclusively "front-end" activity and failing to document changing understandings of the program description as the evaluation unfolds." (F) would certainly be an error. Intervention checks are needed. Cooksy et al. (2001) show the utility of documenting implementation in building a causal argument. From a cultural viewpoint, the add-on might be assuming that the program is implemented appropriately across all relevant dimensions of cultural diversity (e.g., sexual orientation may be key for one program, multiple ethnic heritages and spoken languages for another). "Uniform" implementation may not be the most important criterion (in a one-size-fits-all sense); adaptations may be required to accommodate cultural diversity.

Illustrative Case 1—Description. The evaluand is a secondary school tutorial program. Screening involved teacher judgment and grades. No labels (e.g., "high risk") were attached to students needing help—a plus in this scenario. The program intervention lasted two years, at the end of which time evaluators found no significant differences between the randomly assigned treatment and control groups. Students in both groups had improved but were in need of continued remedial support in the judgment of teachers. Because neither the experimental tutoring nor the control group condition had been documented, evaluators were unaware that the control group had received tutoring from former teachers in the community who had volunteered their time. The intended randomized comparison was invalidated by control group compensation. No cultural information is provided, but clearly this secondary school has a very active and effective PTA (dated language? Most are no longer called this). The PTA, in cooperation with the school principal, mounted the volunteer tutoring of the control group.

Illustrative Case 1—Analysis. My first reaction to this case was, "Would anyone actually *do* such an evaluation any more?" But then a reality check on the revitalized love affair with experimental method (cf. The *What Works Clearinghouse* and other manifestations of Evidenced-Based Practice [EBP]) suggests that Black Box evaluations are still a possibility. Clearly, intervention checks are a necessary and appropriate component of outcome evaluation, and the analysis affirms the need to have monitored implementation over the two-year period of the intervention. The analyst specifically suggests monitoring program activities, time and resources spent as minimum components of documentation. The final paragraph lists a number of data-gathering strategies that would have been potentially useful to consider.

Illustrative Case 2—Description. The setting is a technical training program of an unspecified company. The evaluand is the computer-based training (CBT) phase, one of three components of training. A seven-member task force is established from various divisions of the organization. No cultural context information is mentioned. To familiarize themselves with the evaluand, the task force members participated in a demonstration of CBT. The evaluator also assessed political and organizational constraints by questioning task force members. The evaluation of CBT was expanded based upon these conversations. The final report included a description of CBT, strengths and weaknesses, and recommendations.

Illustrative Case 2—Analysis. The analysis applauds the evaluator for arranging the demonstration and for discussing political and organizational constraints. Both the example and the analysis seem quite dated. Use of technology is now commonplace, and it would be appropriate to describe its use in more detail than appears in this illustrative case.

Neither case really illustrates the complexities of this standard; e.g., the evolving understanding of program descriptions or how different perspectives are synthesized into a single coherent description.

Supporting Documentation.

Cooksy, L. J., Gill, P., & Kelly, P. A. (2001). The program logic model as an integrative framework for a multimethod evaluation. *Evaluation and Program Planning*, *24*, 119-128.

Ridley, C.R., Mendoza, D.W., Kanitz, B.E., Angermeier, L. & Zenk, R. (1994). Cultural sensitivity in multicultural counseling: A perceptual schema model. *Journal of Counseling Psychology*, *41*(2), 125-136.

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

Weiss, C. (1997). How can theory-based evaluation make greater headway? *Evaluation Review, 21*, 501-524.

Return to Table of Contents

A2 Context Analysis.

The context in which the program exists should be examined in enough detail, so that its likely influences on the program can be identified.

Standard. This is an extremely important standard for our attention. It has the (untapped) capacity to address cultural context, so strengthening this standard could be key to improving cultural competence within the current Standards structure. The standard itself is clearly worded, though it presents the influence of context on program as unidirectional when in fact there may be important bi-directional relationships to consider.

◙ This is the standard that directly applies to our concerns and should be the focus of our attention. It is important to note that culture is not included in the definition of context. More important the overview ignores participant characteristics and their impact on accuracy. It requires a major rewrite. Also at least one of the case studies should directly examine an evaluation in which the "participant culture" was drastically different from that of program staff and or the evaluation staff.

Overview. The overview begins by casting an appropriately broad net around context, but cultural context is noticeably missing from the list. In fact, the only dimension singled out is cited as a negative, "impoverished economic conditions." This raises the issue of how context is viewed and the importance of scrutinizing descriptions for subtle (or obvious) bias and prejudice. Here, as in the evaluation itself, both strengths and limitations should be considered.

The overview of this standard is written from an experimental perspective, drawing attention to isolating variables, sorting out causal relationships, establishing representativeness and supporting external validity (paraphrased language in paragraphs one and two). The third paragraph takes the stance of an external evaluator choosing when and where to conduct an evaluation. I think the experimental stance is especially jarring here because it is an unexpected juxtaposition. I always think of this standard in much more qualitative, cultural anthropological terms. It's a good reminder of the importance of reflecting on the assumptions of the evaluator and how that shapes what is noticed and understood, in this case about context.

Guidelines. (A) omits cultural context, which should be explicitly mentioned alongside "technical, social, political, organizational, and economic context." I would also add historical context to the list, a very important component of cultural context but also a noteworthy dimension in its own right. Evaluators need to understand their evaluands in terms of their history and stages of program development, not just as entities frozen in time. The list of illustrative sources should also be expanded beyond archival sources to include living histories, conversations and direct observation (with varying degrees of participation, as appropriate). (B) exudes the mentality of experimental control, logging "history" as threats to internal validity. The idea of journaling is certainly relevant, expanded beyond the tightly specified lens portrayed in B. (C) makes some degree of sense within an experimental paradigm—recording various forms of intervention diffusion or other contamination that would invalidate causal inference concerning experimental and control group differences. But again, this is *way* too narrow a construal to be helpful as a general guideline. Understanding the relationships between individuals and the evaluand is an important context piece that is more complicated than a dichotomy of "interferes with" or giving "special assistance." For example, where do these stakeholders fall on the internal/external continuum of proximity to or distance from the evaluand? What is their history (perhaps intergenerational) with the evaluand? What are the power dynamics that define or underlie these relationships? (D) has broad relevance to cultural context, even though this is not brought out. Though D is written through the lens of external validity, it calls for attention to context that extends beyond causal models.

Overall, this is an extremely relevant and important guideline that has been narrowly operationalized in its current version. **Expanding A2 should be a key focal point of our effort to promote appropriate infusion of cultural diversity into the Standards.** To me, A2 epitomizes the "errors of omission" regarding cultural competence (whereas U2 Evaluator Credibility epitomizes the "errors of commission").

Common Errors. (A) supports our contention that ignoring cultural context is an error. This should be stated explicitly. I personally would also tie this to construct validity (as opposed to the current emphasis on internal/external validity), drawing upon Messick's (1995) notion of construct underrepresentation as a way of framing narrowness. (B) No argument that B is an error. Interestingly, I always find scrutiny of public relations documents to be a particularly interesting window into the values and assumptions of the program with respect to cultural diversity. So I would definitely support a cultural reading of such documents, though not accepting them uncritically as truth. With respect to (C), my concern echoes what I've said before about these "cautionary notes" on time. Because the steps necessary to support multicultural validity are often time consuming, I worry that such cautions could undermine multicultural validity. On the other hand, no one would argue against managing time well and balancing the energies spent on different activities.

This standard is underwritten in its current form. Additional errors surround taking a majority perspective as truth, failing to understand the historical context of the evaluand, and failing to explore the diversity *within* culturally diverse sub-groups.

Illustrative Case 1—Description. The evaluand is a secondary school program introducing students to computers—one of the badly outdated illustrations that need to be replaced. The reader is told the size of the school in terms of both faculty and students, but no other cultural information is given. We are told that the inventor of the program is a mathematics teacher, but nothing else about the origins of the evaluand. The gist of the illustration is that the program was not implemented in a timely way due to contextual factors (new principal, campaign to raise school taxes) but the relevance of these factors to the limited implementation is not explained. The evaluation was performed by a panel member from the program's sponsor, the state education agency. The panelist "was assigned" this project and given two months to complete it. No historical context for this action is provided, nor is procedure by which each panelist was assigned a project to evaluate described. As an illustration of context analysis, this is a very thin case, offering none of the rich detail or deep understanding that one would hope to see accompanying this standard.

Illustrative Case 1—Analysis. The analysis focuses only on the fact that the negative judgment was rendered without considering what are portrayed as mitigating circumstances. The analyst recommends that "a file of contextual information, including reports of conditions and events in the school, school district, and community during the life of the program" should have been kept—all reasonable suggestions. But the politics surrounding this program (e.g., who is the math teacher who designed the program, and what is his/her role and status within the school?) and its evaluation (e.g., who comprises the panel, and who was the panelist assigned to this evaluation? Why only two months to complete the study?) are never explored, nor are logistical issues of context (e.g., when were students to be exposed to this program and how did that fit in the context of other curricular units?) Interestingly, the analyses refers to "the panel" gathering data and reaching a negative judgment, whereas the Description explicitly states that *one* panel *member* was assigned to evaluate this project.

By omitting any cultural context from case illustration and analysis, the standard sends a message that such considerations are unimportant.

Illustrative Case 2—Description. The evaluand is a one-day training program for salespersons in the sales division of a large corporation. The evaluator is external to the corporation. Though the procedural details are sketchy, the description implies that the training (which is on the topic of identifying new customers) occurred in groups of 75 trainees and the evaluation methods included direct observation of training sessions, a paper and pencil questionnaire completed by trainees, and a post-training focus group of twelve trainees per session. No information is provided on the culture of either the corporate organization or the participants. The first session (home office) received a positive evaluation; the second session (district office) was more negatively evaluated. Context enters the picture when the participants in the second session's focus group describe a major reorganization of field personnel, including reassignment of some of the people in the training. The description attributes the negative evaluation of the training to the fact that trainees were disinterested in the content, given their impending reassignments, and that at least some trainees were displeased by prospects of reduced income and status.

Illustrative Case 2—Analysis. Though the reorganization and accompanying reductions in salary and status are indicative of organizational culture and climate, the analysis does not focus on this. Instead the analysis recommends that the evaluator ask the client about "conditions in the company that might affect response to the training sessions," possibly delaying training "until the negative effects of the reorganization had dissipated." The relevance of the training to the reorganization is not explored, nor is the historical context of either. The analyst also makes the fairly predictable recommendation that the evaluator find out who is invited to participate in training and what their expectations are. Cultural diversity of the invited participants is never addressed, though the description is clear that there are differences in income and status; perhaps there are unexplored equity issues within the corporation. Finally, though we are told that the topic of training is identifying new customers, there is no mention of cultural diversity among current or potential future customers, though this would be a relevant marketing consideration.

Again, **this standard is rich with opportunities to explore the importance of cultural context of programs being evaluated, but none is cultivated in the present text.** This omission weakens the ability of the standards to support culturally competent evaluation.

Supporting Documentation.

Delgado, M. (1996). Puerto Rican elders and gerontological research: Avenues for empowerment and participation. *Activities, Adaptation & Aging , 21*(2), 77-89.

Mercier, C. (1997). Participation in stakeholder-based evaluation: A case study. *Evaluation and Program Planning, 20*, 467-475.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Norwood, P., Atkinson, S., Tellez, K., & Saldana, D. (1997). Contextualizing parent education programs in urban schools: The impact on minority parents and students. *Urban Education, 32*, 411-432.

Pumariega, A. (1996). Culturally competent outcome evaluation in systems of care for children's mental health. *Journal of Child and Family Studies, 5*, 389-397.

Return to Table of Contents

A3 Described Purposes and Procedures.

The purposes and procedures of the evaluation should be monitored and described in enough detail, so that they can be identified and assessed.

Standard. This standard supports meta-evaluation by requiring clear description of the purposes and procedures of the evaluation. The reference to *monitoring* the purposes and procedures acknowledges the fact that they may evolve and change as the evaluation progresses. I also like the fact that purposes is plural, setting the stage for exploring overt and covert purposes as well as purposes viewed differently by different stakeholders.

Overview. The opening sentence of the overview immediately reduces the value of the standard by severely limiting "purposes" to *stated* objectives and *intended* uses of results. Apparently unstated (covert) objectives or unintended uses are off limits. "Procedures" are cast as data gathering, organizing, analyzing and reporting. Though the second paragraph acknowledges potential "differences of opinion" regarding purposes and procedures, culture is not mentioned as a source of difference. The evaluator is advised to identify and assess points of agreement and disagreement and to document understanding at the outset. Though this part of the overview has a very preordinate tone to it, consideration is given to emergent understandings in the subsequent paragraph. Again, the recommendation that "independent evaluators monitor, describe, and judge the purposes and procedures of the evaluation" explicitly ties it to meta-evaluation. The last sentence of the second paragraph (p. 137) leaves open the possibility that not all differences of opinion regarding purposes and procedures can be reconciled.

The third paragraph speaks to changing perceptions of purposes and procedures at different stages during the evaluation, calling for "periodic review" to reflect on the appropriateness of the original plans, and the extent to which they were implemented and/or altered. Evaluators are advised to document the final understandings of purposes and procedures so that conclusions are put in proper context. The final paragraph of the overview connects the standard to meta-evaluation, replication, and instruction of future evaluators.

Though this overview is among the better-developed ones in the Standards volume, culture is missing from the picture of what should be monitored and documented. Differences of opinion regarding purposes may break along cultural lines. Procedures of data collection, analysis and reporting need to select and handle cultural variables responsibly (and meta-evaluation must examine the extent to which this has been done). And cultural context would necessarily be taken into account in planning replications. The overview does not address the importance of documenting unintended uses of results, nor does it address the documentation of process use. This is a reflection of the datedness of Standards; conceptualization of evaluation use has continued to evolve in the decade since the last publication.

Guidelines. (A) While there is nothing wrong with recording a client's conception of purpose and of intended results-based use/influence, this guideline ignores process-based influence, unintended influence (either results-based or process-based), and stakeholder conceptions beyond those of the client. (B) is similarly client-focused, not alluding to other key stakeholders whose understandings should also be thoroughly discussed and recorded. (C) should cross-reference P2 Formal Agreements, as should (E). (D) and (F) are potentially important means to capture cultural components of implementation and/or accommodations required (e.g., translation issues). (G) raises interesting issues of ownership and liability inherent in record keeping; A7 Systematic Information Control should be cited. What procedures should be followed to document and explore covert agendas, for example? Standards of evaluation may meld with standards for investigative reporting. (H) is relevant to evaluations that call for written final reports, but should a technical report always be required? If a client does not request (and does not wish to pay for!) a technical report, what is the evaluator's ethical obligation with respect to documentation? Cultural standards may vary on this point, as previously discussed under U5 Report Clarity, which is appropriately cited. (I) underscores the relevance of meta-evaluation by independent evaluators. While, I agree that this is potentially very valuable when resources permit, I think meta-evaluation should not be equated with only external review. The kinds of documentation referred to in (D) and (F) also facilitate reflective practice and internal meta-evaluation.

◙ Note: The guidelines focus on the "client," ignoring the participants, be they staff or non-employees. The illustrative case references this, but could be more explicit.

Common Errors. (A) and (B). Certainly, it would be an error to assume that the evaluation described in an initial proposal or contract was identical to the completed evaluation without verification. (C) could incorporate failure to adjust the data collection, analysis or reporting strategies to make them culturally congruent, thereby weakening validity. (D) goes beyond documentation to address the "soundness" of purposes and procedures. Several of the Accuracy standards could be cited in support of "soundness," notably A4 Defensible Information Sources, A5 Valid Information, and A6 Reliable Information.

One really interesting facet of this standard is that it begins with the importance of documenting data gathering and moves forward from there. What's missing is the link between the alleged purposes of the evaluation and the evaluation questions. To me, another error is failing to scrutinize, describe, and document how the purpose of the evaluation is translated into evaluation questions.

Illustrative Case—Description. The evaluand is a high school mathematics program, for which an independent study approach and a traditional instructional approach are being compared. No information is given concerning the cultural context of the school or district. A formal written agreement between the assistant superintendent and the district evaluation staff specified the purpose of the study and design components such as sampling frame, random assignment of teachers to instructional approaches, and outcome variables of interest. The evaluator assumed that the agreement was trustworthy and proceeded to collect and analyze data. Unbeknownst to the evaluator, the assistant superintendent narrowed the purpose of the study and eliminated the random assignment of teachers. This invalidated the conclusions of the study and created a mismatch between the assistant superintendent's new questions and the answers the study was designed to provide. Teachers were quick to point out validity threats.

Illustrative Case—Analysis. This analysis is interesting (read, overtly sexist) insofar as the assistant superintendent (a man) violated the original agreement and compromised the study, but the analyst blames the evaluator (a woman): "The difficulty arose because the evaluator should not have assumed that purposes and procedures, once agreed upon, would remain constant throughout a project." I would seriously scrutinize the dynamics of power and gender in this situation. (No information is given concerning other diversity characteristics that may be in play.) The analyst is clear that the assignment of teachers and students to instructional approaches was not under the evaluator's control, but nevertheless argues that she should have "taken steps to monitor and record changes," meeting periodically with the assistant superintendent to review implementation, methods of instruction and data collection. The trustworthiness of the assistant superintendent is never questioned nor are the ethics of his violating a formal agreement (P2).

In an unrelated criticism, the analyst notes in closing that "The evaluator should have described any differences among stakeholders in perceptions of the purposes and procedures of the evaluation." This is not a poor suggestion, but interestingly no stakeholders beyond the assistant superintendent as client and the math teachers as program providers are mentioned in the case description. Certainly it would have been appropriate to consider the students, their parents, teachers of subjects other than mathematics, and other potentially relevant stakeholders in describing and documenting differences in perceptions of evaluation purpose and procedures.

This is a potentially useful standard to support cultural competence; however, it is narrowly operationalized and the case illustration/analysis is flawed.

Supporting Documentation.

Besharvo, D., & Gardiner, K. (1997). Sex education and abstinence: Programs and evaluation. *Children and Youth Services Review, 19*, 501-506.

Galvin, P. (1999). The politics of research on educational productivity. *Educational Policy, 13* (1), 136-151.

Return to Table of Contents

A4 Defensible Information Sources.

The sources of information used in a program evaluation should be described in enough detail, so that the adequacy of the information can be assessed.

Standard. This standard is a cornerstone of multicultural validity. By calling for scrutiny of the adequacy of information sources, the standard creates a framework for examining the balance of majority/minority viewpoints in an evaluation. “Sources” is appropriately plural, directing attention to triangulation of information sources.

Overview. The overview opens by reiterating the value of multiple sources and providing a limited (and somewhat dated) list of examples. ◙ Add the following after the second sentence of the overview, “Using a variety of data sources will more accurately capture the depth and diversity of the program and its participants.” ◙ In illustrating the importance of triangulation, the overview also balances quantitative with qualitative data, which is a plus. The third paragraph diverts from information sources to information-gathering strategies. While triangulation of strategies is also important, in my experience it is helpful to keep the two issues somewhat separate. Scrutinizing sources is particularly helpful in identifying whose voices are represented and not represented in an evaluation. A premature shift of attention to method often obscures the issue of voice or perspective. The fourth paragraph takes an even sharper detour, this time into sampling, giving a very elementary introductory statement followed by a gratuitously sexist example (“man on the street interviews”). The fifth paragraph returns to information sources with a strong statement that evaluators “should document, justify and report their sources of information, [and] the criteria used to select them. . . ” but then folds data collection strategies back into the same sentence. Attention to and documentation of data collection strategies is certainly important, but they should not be conflated with information sources. This is a common error that should not be perpetuated here. The sentence ends with an interesting caution to document, justify and report “any unique and biasing features of the obtained information.” Cultural bias is not explicitly named, but it could certainly fall within this directive. Certainly cultural diversity and dynamics of power and privilege create “unique features” that should be noted when scrutinizing the adequacy of the information base to answer the evaluation questions posed. *Culturally* defensible information sources should be part of the conversation surrounding evaluation credibility, but I agree with the need to provide adequate detail in description and documentation of sources so that this meta-evaluative reflection can occur. The overview concludes with an oddly condescending assertion that “careless or uninformed stakeholders” could be misled by a report in which the information sources were inadequately documented. In fact, *anyone* could be misled by such a report; it should reflect negatively on the evaluation rather than the stakeholders.

This overview is poorly written, plain and simple. It is choppy, lacks clear logic development, and repeatedly strays off the topic of the standard.

Guidelines. The guidelines continue to conflate information sources, sampling frames, and data gathering strategies. (A) speaks only to defining a population and describing sampling procedures. Similarly, (D) addresses the need to document changes in the sampling frame. In (B) a clear statement concerning sources of information is followed by a list of examples that mix information sources with data gathering strategies. (C) is only about information gathering strategies, not about information sources. (E) refers to decision rules involved in data collection. (F) is appropriate to the use of archival data as an information source. Here, “soundness” seems to be a synonym for validity; A5 valid Information should be cited. **(G) is the crux of the standard as I understand it** (and as it is presently written). “Limitations” should include the extent to which the information is culturally bounded.

This standard is written to address information sources, but the narrative pulls in issues of sampling and data collection. I would find it clearer if these matters were pulled apart into separate standards. That would create the space to address the unique ways in which cultural diversity needs to be considered in information sources, sampling, and data collection respectively, with sufficient specificity to be useful. If the intention is to address all of these here, then the standard itself should be rewritten to make that clear. *Common Errors*. (A) correctly cites it as an error to provide insufficient detail. The source here is unclear, however, mixing source and data collection strategy. If one is compiling student portfolios specifically for the evaluation, then the source (whose voice?) is the student and the information-gathering strategy is portfolio assessment. If the portfolios already exist, then the source is archival records and the strategy involves sampling, coding and content analysis of the records. (B) raises issues of reliability and validity of information sources; A5 Valid Information and A6 Reliable Information should be cited. In these examples, the sources are project staff and evaluation staff, with the data gathering strategy written reports. (C) is particularly relevant to cultural diversity—noticing what is most valued and what is less valued or discarded. I especially support the recognition that any single information source is necessarily imperfect and limited. This can be used to examine alternatives to majority sources and foster the inclusion of more diverse perspectives (which, in turn, have their own limitations to be noticed). Building on Madison's notion of primary inclusion of consumers and Scriven's emphasis on consumers in the evaluation process, I always try to notice where the consumers of programs have been included or omitted as information sources (as well as how well triangulated the strategies are for tapping that perspective). (D) Certainly this is an error worth mentioning, although I think the field has moved beyond the stereotypes listed here. Both qualitative and quantitative data are subject to measurement error (A6 Reliable Information) and distortion (A5 Valid Information). Moreover, the justifications for validity may vary across cultural dimensions. What is considered credible data to one audience may be disregarded by another. **The fact that what is considered to be a "defensible information source"—and how that judgment is "defended"—varies across stakeholders and may be culturally bounded is missing from this standard.** (E) puzzles me. I think the intent of it is the same as raised in previous standards—a caution against over-allegiance to the standard at the expense of other important activities or concerns. But here I don't quite grasp what this would look like: obscuring or overlooking the *content* of the information while documenting its *source*? It just seems like an odd juxtaposition to me. (F) is a technical point, calling for power analysis in determining sample size. *If* one is going into sampling issues, there are many issues related to cultural diversity that should be considered. As I've said before, I personally feel that sampling should not be folded in with information sources; it should have its own standard, permitting a more appropriate level of detail in the discussion as well as creating room to address the proper treatment of diversity variables in sampling. *Illustrative Case—Description*. The evaluand is a teacher education program to prepare for instruction in open classrooms (another dated example). The Chair of the Board of Education at the state level called for the evaluation, and the evaluator was a school district principal. No cultural context information is given about the program, the district, or the evaluation, but we are told that the program was "controversial." The information sources were: superintendents, principals, students in the program, graduates of the program, the program director and her staff. Data collection strategies included questionnaires and interviews. Program critics attacked the evaluation based upon methodology (sampling bias and low response rates). Funding was cut and the program was subsequently terminated.

While this is a pretty good illustration of F2 Political Viability, it really does little to explore the information sources themselves. In fact, there is only one sentence specifically on information sources in the whole two paragraphs of the Description, and that is a simple source list. No information is provided that would allow the reader to reflect on whether the choice of sources was defensible or whose perspectives might have been omitted. The potential of this standard to examine cultural perspective is never realized.

Illustrative Case—Analysis. The entire analysis focuses on sampling strategies, including analysis of respondent/non-respondent bias. There is no conversation about the information sources themselves. "Defensible" is operationalized as statistically representative. The politically-motivated nature of the critique described in this case is not addressed, nor is there any discussion of what information was more or less valued and by whom. To me, this case illustration and especially its accompanying analysis pretty much misses the whole point of this standard.

Supporting Documentation. ◙ I have cited Madison and Scriven several times already, and their work is very influential in my thinking.

Hoefer, R. (2000). Accountability in action: Program evaluation in nonprofit human service agencies. *Nonprofit Management & Leadership, 11*, 167-177.

Wolfer, T., & Johnson, M. (2003). Re-evaluating student evaluation of teaching: The Teaching Evaluation Form. *Journal of Social Work Education, 39*(1), 111-121.

Return to Table of Contents

A5 Valid Information.

The information gathering procedures should be chosen or developed and then implemented so that they will assure that the interpretation arrived at is valid for the intended use.

Standard. To me, this is the crux of meta-evaluation and of cultural critique; it's all about validity. Unfortunately, the standard is written to focus narrowly on measurement validity, specifically, the choice/development of measurement tools and the implementation of data gathering procedures. This is too limited a perspective on validity. On the positive side, validity is tied correctly tied to use, though here again, I would broaden the conversation to include unintended as well as intended uses. Unintended uses may or may not be valid, depending upon the specifics of context.

◙ Rewrite: "The information gathering procedures utilized should be those that assure a valid interpretation of the findings within the context of the evaluation and for their intended use." (or something to this effect.)

Overview. The overview opens with a traditional definition of validity, narrower than Messick's unified validity theory, which includes actions as well as inferences. This is not surprising, given the conservative stance of contributing evaluators on this point.

Shadish and I have debated this point on more than one occasion, and while some of Messick's language is reflected in the current *Standards for Educational and Psychological Testing* (4^{th} ed.), which should certainly be referenced here, Messick's untimely death stalled efforts to broaden the construct of validity. We should at least argue for the language to be updated to be congruent with that of the 4^{th} edition *Testing* standards; that will broaden is a bit.

The standard proceeds to present an overview of elements of test validation. I agree with Cronbach's stance that validation *is* evaluation, so much of this becomes meta-evaluative (A12). Culture is not addressed, though it saturates every element in this process: the definition of the constructs; the types of information to be collected (from whom and how); the procedural steps followed; the data synthesis, scoring and interpretation; and choice of particular justifications to support validity. This is not a particularly radical notion. Even in the current description, the overview notes (p. 146), "the validity of an inference depends upon the evaluation questions being addressed, the procedure used, the conditions of data collection, the judging and scoring procedures followed, the analysis procedures used, and the characteristics of persons who provided the data or information." To the latter, I would add, reflexively, "the characteristics of persons who *collected* the data or information and who *designed* the study in all of the ways previously noted." The framework is in place for examining the interaction of culture and validity, though this topic is not introduced. I would argue that validity requires cultural competence of the evaluator(s), appropriate to the context and topic of study; therefore, validity cannot be responsibly considered apart from culture.

Critical multiplism is alluded to in the discussion of triangulation of variables, information sources, and measurement procedures (p. 146). (Note that the importance of disentangling information sources from measurement procedures is underscored here; one has to consider them separately to enter into conversations of post-positivist critical multiplism.) The closing paragraph of the Overview slips into the information source-as-method confusion again, although the point concerning the fallibility of any one source (or method) is well-taken. The culturally-bound nature of inferences is not discussed, though it could easily be brought into the conversation.

This Overview is richer in detail than many of the others; hence there is more to critique but also more room to revise and insert cultural considerations. Clearly, entire texts have been written on validity, so there is a limit to how thorough coverage can be in one summary. Still, I am optimistic that if the source documentation were updated and cultural context brought into the discussion, this standard could support multicultural validity (without necessarily introducing the term, by the way. The language within validity theory is quite dense, so I agree with the author's stance in not introducing distinctions within the general construct.)

◙ General comment: This standard and all of the guidelines should be revised to incorporate issues related to the validity of different data gathering methods and instruments for diverse populations and the extent to which interpretations could be misleading if the evaluator (or evaluation team) does not understand the potentially influence of contextual factors.

Guidelines. (A) is a helpful guideline to check program congruence; a similar procedure is advisable to check cultural congruence. The door to this is opened when the guideline notes that the check should be informed by "representatives of important stakeholder groups" as well as by program providers. The cultural diversity among direct and indirect consumers of the program could be tapped, for example, to bring to the evaluator's attention dimensions of culture that should be attended to and respected in the evaluation design. (B) speaks to the importance of documenting justifications of validity, a guideline I wholeheartedly support. Particular attention should be given to describing cultural context (A2 Context Analysis) and documenting the evidence supporting the appropriateness of the procedures for use in this context. (C) Definitely important to cite the Standards, though I would propose a caveat indicating that they are not the only relevant source document on validity. (D) In contrast to the other guidelines, this one seems to be a fairly specific point regarding the types of evidence that may support validity. (E) Truth telling is not just a matter of personal motivation, as this guideline implies, but of systemic issues of power and contingencies (actual or perceived) attached to the data being gathered. Certainly, at the micro level, one should develop and pilot instruments to achieve clear, culturally-appropriate expression and minimize confusion or bias, but at the macro level, it is also important to consider the ways in which data are likely to be used (and by whom), and how these uses may undermine or support truth telling. Validity presumes a certain level of trust, predicated at least in part on the assumption that risks have been anticipated and guarded against (P3 Rights of Human Subjects). In evaluation, there may be very real risks (e.g., program dissolution, termination of services) that cannot be eliminated, as well as systemic biases (e.g., racism or classism) that engender mistrust. From this perspective, validity threats are woven into the fabric of programs and their social contexts. Examining validity requires a very wide-angle lens. The lens here is pretty narrow. ◙ E: whose truth are we talking about? ◙ (F) could explicitly mention cultural competence as one way in which evaluators should be "qualified and adequately prepared." ◙ F: add training, practice "and sensitivity to the context of the evaluation." ◙ (G) begins to address this when it states that "proper account must be taken of context (see A2, Context Analysis), the characteristics of the subjects or groups with whom the procedure was used, and the qualifications and training, if needed, of the individuals who administered or used the procedure." **This phrase comes as close to addressing our concerns as any standard in the book.** Though the application is still focused narrowly on measurement and the wording is dated, it is a step in the right direction. (H) speaks to the use of archival data, unobtrusive measures, and sampling, all of which could have cultural dimensions. Though the authors cite F3, Cost Effectiveness, I would connect it with F1, Practical Procedures. (I) is a meta-evaluation (A12) issue that is extremely important. Validity extends beyond individual measurement tools or procedures to examine the integrity of the entire evaluation design. Whether the information is sufficient to answer the questions posed is an important piece of this puzzle. (J) explicitly cites the diversity characteristics of "reading ability, language proficiency, or physical handicaps" as illustrations of respondent characteristics. "Physical handicaps" should be replaced by "disability," and the list should be broadened to include potential influences of gender, age, race, ethnicity, and social class at a minimum. Depending on the program being evaluated, religion and sexual orientation may be relevant considerations. (K) is an odd reference to procedures for analyzing qualitative data. I guess the link to validity comes from the extraction of meaning from open-ended responses. While I don't disagree with the guideline, it seems too superficial to stand alone and is perhaps more appropriate under A9 Analysis of Qualitative Information.

This standard, while making only limited explicit references to culture, is one of the better ones in terms of level of detail and development. The Overview and Guidelines actually extend the scope of the standard itself, and the link between culture and validity is outlined if not fleshed out.

Common Errors. (A) should also address the need to consider the cultural congruence of the instrument in the proposed application context. (B) Mono-method bias is a well-recognized threat to internal validity, another example of how the text discussion of this standard is broader than the standard itself. (C) Amen to that one! Especially but certainly not exclusively in matters of cultural congruence, validation is complex and time consuming. (D) But the caveat here is taking existing procedures developed and validated on majority populations and applying them in minority settings without further validation. That is a potentially even more serious error than ignoring existing instruments. (E) Yes, and a corollary error here is failing to consider the cultural competence of persons who will be working with, or gathering information from, persons different from themselves. (F) Yes, and by "adequately conducted," this should include checking that observations were done in a respectful manner, congruent with the norms of the setting, and with appropriate consent of persons being observed. (G) Certainly, it is important to have instruments reviewed by relevant stakeholders. The phrase, "allow qualified stakeholders the opportunity" takes me aback, however. To me it suggests a top-down view of stakeholder participation, implying that only a select few stakeholders could meet the qualifications necessary to be granted permission to review instruments. In fact, the reviewers are doing the researchers a favor, not the other way around. The instruments should be reviewed by persons who are as similar as possible to those with whom the instruments or procedures will be used. I agree that it is an error if evaluators fail to build this into their procedures. (H) This error is the flip side of guideline (J) above, interestingly with the same diversity dimensions used as illustration: reading ability, language proficiency, and physical handicap. Failing to consider the characteristics of respondents that affect their reactions to evaluation is indeed an error, but the lens of relevant characteristics needs to be expanded. Not only is rote repetition of categories simplistic, it also implies that only "special" or "handicapping" conditions need be considered.

Illustrative Case 1—Description. The setting is a middle school, and the evaluand is a set of curriculum units on ecology. No information on cultural context is given. The curriculum committee of the middle school requested the evaluation, and the evaluators were from the school district. The curriculum committee wanted to know if knowledge of environmental issues such as conservation and endangered species had improved and whether student behavior such as littering had changed. Inexplicably, the evaluators chose a questionnaire on school citizenship and portions of a national standardized test on hygiene, biology, and earth science, and administered them within a pretest/posttest design. No differences were found, a predictable finding given the irrelevance of the instruments to the curriculum units.

Illustrative Case 1—Analysis. The case analysis points out the obvious disconnect between the evaluation questions posed and the content of the data gathered—zero content validity of these measures for this application. Importantly, the analysis also notes that direct observation of the program would have given evaluators an opportunity to notice unintended outcomes. Creative ways of monitoring changes in student behavior through unobtrusive measures and direct observation are also mentioned, all of which have high face validity.

This is a pretty straightforward example of failed content validity. Given the complexity of the topic, a more challenging example would be more useful. I do like the fact that the illustration avoids statistical jargon and symbols that could make the example less accessible to some readers, however.

Illustrative Case 2—Description. This case introduces economic status as a diversity dimension, but no other participant characteristics and no staff characteristics are described. (The term “economically disadvantaged populations” should be replaced by language that emphasizes persons, not a label.) The evaluand is collaboration activities among agencies coordinating education and training for persons living in poverty through a variety of interagency projects. Information on coordination-enhancing activities was collected by a combination of interview, direct observation, and documents review. The evaluators began with a review of the literature to identify core features of interorganizational collaboration, followed by applying the model to a sample of prior case studies, stakeholder rating of core features in focus groups, and blind coding of interview segments using the coding scheme. The coding scheme was revised after each iteration, and the final codes were applied to each interagency project. Local projects were given the opportunity to nominate additional codes. Four local codes were “identified and corroborated.”

Interestingly, the case description mentions in the opening paragraph that the that the interviews (presumably referring to the focus group interviews) were of project staff *and participants*, but the participant voice, which introduces economic status and potentially rich additional diversity dimensions, is not visible in the subsequent procedural description. The focus groups are characterized as generating remarkably consistent major themes; presumably staff and participants were of one mind? Or perhaps majority opinion held sway? Without more detail regarding dimensions of difference and how the staff/participant power differential was handled in the focus groups, it is difficult to assure multicultural validity here.

Illustrative Case 2—Analysis. The analyst applauds the evaluators for using literature to generate code categories and validating those categories through the focus groups. The analyst cites the theoretical meaningfulness of the code categories alongside meaningfulness to local stakeholders and provision for emergent categories during data collection as supporting justifications for validity. Absent any information on cultural context, it is difficult to put the analysis in perspective. Certainly these evaluators were scholarly and thorough, but the question remains has to what dimensions of diversity were relevant in each of these local settings and to what extent the code categories attended to those dimensions. Theoretical models derived from majority literature may be culturally incongruent with local programs. One would hope that these evaluators were attentive to these dimensions in validating their coding scheme. Based on the information given, this remains an open question.

Supporting Documentation. ◙ Validity is a topic that generates heated debate and about which there are strongly held beliefs. The supporting documentation tilts toward the post-positivist perspective. Here we should argue for the inclusion of authors who have explicitly addressed cultural validity and diversity. Feminists are especially strong here. Patti Lather. Sandra Harding. Donna Haraway. I'll include a few sources here; the list could go on and on.

Alkon, A., Tschann, J., Ruane, S., Wolff, M., & Hittner, A. (2001). A violence-prevention and evaluation project with ethnically diverse populations. *American Journal of Preventive Medicine, 20*, 48-55.

Kirkhart, K. E. (1995). Seeking multicultural validity: A postcard from the road. *Evaluation Practice*, *16*(1), 1-12.

Lather, P. (1986). Issues of validity in openly ideological research: Between a rock and a soft place. *Interchange*, *17*(4), 63-84.

Lather, P. (1993). Fertile obsession: Validity after poststructuralism. *The Sociological Quarterly*, *34*(4), 673-693.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Mishler, E. G. (1990). Validation in inquiry-guided research: The role of exemplars in narrative studies, *Harvard Educational Review*, *60*(4), 415-442.

Psacharopoulous, G. (1995). Using evaluation indicators to track the performance of education programs. In R. Picciotto & R. C. Rist (Eds.), *Evaluating country development policies and programs: New approaches for a new agenda, New Directions for program Evaluation,* No. 67 (pp. 93-104). San Francisco: Jossey-Bass.

Shadish, W. R., Jr., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Waldofgel, J. (2000). Child welfare research: How adequate are the data? *Children and Youth Services Review, 22*, 705-741.

Return to Table of Contents

A6 Reliable Information.

The information gathering procedures should be chosen or developed and then implemented so that they will assure that the information obtained is sufficiently reliable for the intended use.

Standard. This standard refers specifically to measurement reliability. This is relevant to multicultural validity, since methodological justifications of validity focusing on measurement require such reliability. This is only the tip of the iceberg here, however. The Guidelines—specifically (D) and (E)—expand the scope of the standard to address concerns of evaluator and stakeholder perspectives more clearly than many of the other standards, encouraging the kind of reflection that is a necessary component of cultural competence.

◙ Rewrite: "... procedures utilized should assure that the information obtained is reliable for the intended population and use."

Overview. The overview gives a general, non-technical definition of reliability, correctly distinguishing between random error and systematic sources of variability. "Characteristics of program participants" is cited as an example of a systematic source of influence. The overview conversationally summarizes different types of reliability without slipping into jargon, but noting that "different information-gathering procedures are sensitive to different sources of this random error." Two important points are raised at the end of the overview that are relevant (albeit somewhat indirectly) to the treatment of cultural diversity in measurement reliability. First is the unit of analysis. The reader is reminded that if the unit of analysis is a group, then it is inappropriate to assess reliability at the individual level. Second, age diversity is mentioned when the overview notes that "reliability estimates derived from procedures used with adults cannot automatically be extended to adolescents." The same principle applies to other dimensions of difference.

◙ This standard could benefit from specific references to cultural competence.

Guidelines. Though culture is not explicitly mentioned in (A), this is a particularly relevant guideline. The caution against assuming that previously favorable reliability results can be generalized to other groups or information gathering procedures encompasses generalization across dimensions of cultural difference. Majority results ought not be generalized to minority settings; instead information should be collected that is directly relevant to the use of instruments or procedures in those settings. ◙ Guideline A, add the following example after the first sentence, (e.g., generalizing procedures across culturally diverse contexts may be inappropriate.) ◙ (B) explicitly mentioned the "heterogeneity of persons in terms of the characteristics being measured or observed" as a factor influencing reliability. This is relevant to groups that may have been selected (e.g., for educational or social programs) because of high or low scores on a variable of interest. Such groupings may break along cultural lines as well. (C) is definitely relevant; the source references need to be updated, however. **(D) is a gem tucked in among more technical considerations of measurement error. It is one of the few places in the Standards that any attention is given to the evaluator's reflection on his/her own "posture and values and their [*sic*] role in the inquiry." The presumed ability of a peer to be "impartial" may need further examination, but this guideline is definitely headed in the right direction! In the sprit of taking a strengths perspective with the Joint Committee, this can be cited as a positive! (E) continues the conversation, calling for monitoring evaluator expectations as a check on "a predominant influence of the evaluator's own perspective." The call for maintaining "sensitivity to the perspectives of the stakeholders" fits perfectly with cultural sensitivity, and the explicit mention of "alternate explanations for the phenomenon observed" opens the door to examine the culturally-bound nature of "propositions, interpretations and conclusions mentioned" in (D).** (F) returns to more traditional guidelines for insuring inter-rater reliability of scoring, categorization, and coding. The mention of an outside auditor should bring in A12 Meta-evaluation as a relevant cross listing. (G), while generically framed, could include training that sensitizes data analysts to mistakes likely to occur across dimensions of cultural difference.

Common Errors. I agree that (A) is a common error. I struggle to get my students to appreciate that the various types of reliability are not interchangeable. It seems appropriate to open with this fundamental distinction. (B) is highly relevant to cultural diversity. The "differences between the setting and sample of the reported reliability study and those of the evaluation" are often differences in cultural characteristics. (C) is a general principle that applies across dimensions of cultural difference under the categories of "how, when, and to whom" an instrument is administered. (D), (E), and (F) focus on technical considerations. (D) is a statistical point worth remembering, relating back to the unit of analysis issue, while (E) cautions against confusing the reliability of dichotomous judgments with the reliability of the continuous data on which the judgment is based. (F) is a statistical note concerning the reliability of difference scores. (G) through (J) speak to the relation between reliability and validity (A5). Outcome data containing high amounts of measurement error should not be considered persuasive (G). Though validity requires reliability, reliability does not insure validity (H). A measure could be consistently irrelevant. (I) reminds us that it is indeed an error to assume that an evaluator can step outside of his/her perspective, training or previous experience in making observations or rendering judgments. **To me, this is an important caution, directly relevant to an evaluator's inability to step outside of his/her own cultural context and experience.** (F) speaks to the importance of considering all relevant information in making inferences and interpretations. I would cite U3 Information Scope and Selection and A10 Justified Conclusions here.

Illustrative Case 1—Description. The evaluand is an innovative instructional technique, based upon the sample, presumably one used in elementary classrooms. The evaluators are not identified. Measurement tools to be developed were an observation checklist and several objectives-referenced tests. Each test addressed ten objectives and contained five items per objective. There were two forms of each test, and each form was given to a different sample of second and fourth graders at six different schools. No information is given on cultural context of the schools or diversity of students. Internal consistent of each form in each application exceeded .80. The observation form is not described, but we are told that it was piloted by a single observer in each of the six sites and that this observer studied a single domain on two occasions, three weeks apart. Correlations between observations were calculated at the individual item level, and all exceeded .60. Based upon these data, the evaluators deemed the measures sufficiently reliable to proceed with the evaluation. At some unspecified later point in time, a group of teachers challenged the reliability of the tools, pointing out that alternate forms reliability had not been establish nor had inter-rater reliability at a single site. They also pointed out that while internal consistency of the tests had been examined for the total score, no analysis of subscale scores by objective had been conducted, despite the fact that it was these subscales scores that were primarily used by teachers to assess student mastery. The context in which these challenges were raised is not explained.

Illustrative Case 1—Analysis. The analysis points out that internal consistency and test-retest reliability are not always the most appropriate types of reliability to consider. With respect to the tests, alternate forms reliability could have been established by administering both versions to the same students. (The matter of timing is a little unclear and order effects are not mentioned.) Since pass-fail judgments are to be made based upon these tests, the consistency of these judgments, not just of the scores themselves, should be established. With respect to the checklists, simultaneous independent observations by two raters observing the same classroom could be compared to establish inter-rater reliability. Since no context information is provided, the analysis cannot comment on matters of perspective, cultural or otherwise.

Illustrative Case 2—Description. The setting is a medical school, and the evaluand is a course on fundamentals of clinical medicine. A radiologist wanted to evaluate the extent to which the course contained instruction in the use of diagnostic imaging procedures. He designed an observational rating form and hired two first-semester medical students to conduct classroom observations. The form list four types of diagnostic images, but did not define them. Observers were to record frequency counts by type and also record content area of application. Armed with these forms and a lecture schedule, the two observers made independent observations. (Note: it is unclear whether they independently observed the same lecture or they divided the lectures between them.) The radiologist's secretary clerk tallied the data, and the radiologist drew conclusions and submitted a report to the curriculum committee, calling for change. Course instructors challenged the findings, and subsequent scrutiny by the committee revealed inconsistency and confusion between the observers about the meaning of what was observed. The embarrassed radiologist retracted his report. No information is presented on cultural diversity of the setting, the observers, the instructors, the investigator, or members of the curriculum committee.

Illustrative Case 2—Analysis. The analysis points to three errors committed by the radiologist: failure to train the observers carefully and to check their inter-rater reliability during a pilot; failure to monitor the observations during the actual study and spot-check inter-rater reliability himself; and failure to discuss his findings with a colleague who could have saved him public embarrassment by pointing out errors. The latter raises an interesting meta-evaluation (A12) point that extends beyond reliability, but the recommendations to establish inter-rater reliability are well taken. Given the apparent confidence with which the radiologist issued a strongly worded report based upon weak data, I am driven to curiosity about his posture in this evaluation. Did he enter into the study with a particular opinion he sought to support? Did he lack this most rudimentary research knowledge, or had he forgotten it over time, or was he so convinced of his own position that he arrogantly ignored basic procedures? Despite the fact that this standard clearly draws our attention to such matters, the analyst makes no comment. This analysis misses an opportunity to examine the influence of evaluator perspective, training, or previous experience.

Supporting Documentation. ◙ Though in need of updating, the source citations include a balance of qualitative and quantitative perspectives. The link to evaluator posture and sensitivity to alternate perspectives opens up a wide range of possible source references.

Bogo, M., Regehr, C., Hughes, J., Power, R., & Globerman, J. (2002). Evaluating a measure of student field performance in direct service: Testing reliability and validity of explicit criteria. *Journal of Social Work Education, 38*, 385-401.

Richard, A., Bell, D., Elwood, W., & Dayton-Shotts, C. (1996). Outreach and program evaluation: Some measurement issues. *Evaluation Practice, 17*, 237-250.

Ridley, C.R., Mendoza, D.W., Kanitz, B.E., Angermeier, L. & Zenk, R. (1994). Cultural sensitivity in multicultural counseling: A perceptual schema model. *Journal of Counseling Psychology*, *41*(2), 125-136.

Return to Table of Contents

A7 Systematic Information Control.

The information collected, processed, and reported in an evaluation should be systematically reviewed and any errors found should be corrected.

Standard. This standard, originally titled Systematic Data Control in the first edition, contains a distracting typo in the title that makes the content hard to identify; the title appears in the book as Systematic Information. The standard itself makes it clear that it is the *control* that is to be systematic, not the information. Control refers to review of information to detect and correct errors. This process should be systematically undertaken throughout the data collection, processing and reporting phases of an evaluation. Monitoring attention to cultural diversity could well fall within this standard, though it is not developed in this way.

Overview. The overview begins with a clear statement that the intended purpose of systematic information control is to maintain data accuracy and security. The second paragraph enumerates avoidable errors that should be ferreted out and eliminated. The only item on this list that I would question is, "Information may be collected from an unintended group of respondents." While this may be cast as an error in a preordinate design, it may actually strengthen a design under an emergent paradigm. One might uncover a stakeholder audience whose perspectives had been overlooked or under-sampled and alter the design to add these previously unintended respondents. An example of this might be better portraying the diversity *within* a cultural subgroup by adding respondents.

The third paragraph cautions against a false sense of security in the performance of "even highly qualified and dedicated persons." I agree with the sentiment but I am puzzled by the cross-reference to P1 Service Orientation. That doesn't seem to be a relevant connection, and it risks the false implication that persons dedicated to serving consumers are likely to be less conscientious about monitoring and quality control. The overview addresses training, data security, data control, and accuracy checks to avoid tampering and maximize completeness and accuracy of data. It doesn't explicitly address errors of perspective that could also affect the accuracy or completeness of the data, though it does admonish evaluators to "assess the probable effects of the errors that are not detected." This clause could be invoked to support a quality control review of method to insure that both majority and minority standpoints (on dimensions relevant to the study in question) were accurately addressed and that the information-gathering strategies intended to tap contrasting perspectives had been carried out with equal rigor and thoroughness.

Guidelines. (A) could easily encompass cultural competence training and sensitization to the kinds of mistakes that are likely to arise from cultural insensitivity. This could be spelled out in the guideline or illustrated via a case description and analysis. (B) addresses the need for systematic error checking throughout data collection, data processing, and data reporting, all of which are subject to errors associated with cultural standpoint. The quality control plan should specifically address potential cultural bias. (C) through (F) seem to address technical points surrounding data entry and storage, controlling access and monitoring sub-contractors. The only connection to cultural contexts that immediately comes to mind might be around access and ownership issues. Protecting the integrity of data might mean one thing to federally funded researchers, for example, and something quite different to tribal elders. **(G) is particularly important to**

support multicultural validity. It is good that this guideline emphasizes the need to routinize such procedures and to allocate time to do so.

Common Errors. (A) suggests an interesting caveat, which may be “assuming that no deviation from standardized procedures was necessary to collect data respectfully and accurately.” In other words, while it may be an error to deviate from instructions, it could also be an error to follow instructions mindlessly in contexts in which they were culturally incongruent. (B) covers a wide range of potential problems: errors in the expressive language of the data gathering person or device (e.g., vocabulary, jargon); educational/literacy level of the respondents; spoken and/or written language of greatest fluency; issues of power or trust that may shape the nature of the exchange and what information is shared. To shape this in terms of respondents’ ability to “read , understand, and follow directions they are given” seems oversimplified and condescending. As written, it seems to position the respondent as a passive participant in a process directed by the evaluator. (C) through (F) address technical points concerning scoring, data entry and cleaning, and data management. (G) addresses the skills and experience needed to analyze and report data competently, and cultural competence would certainly come into play here. Staff or consultants lacking culturally-relevant knowledge, skills and experience would be unlikely to recognize errors born of majority perspective. (H) Certainly failing to retract or correct inaccurate results is an error. So too would be failing to present a minority rejoinder to a majority interpretation. (I) is an important reminder not to skimp on quality control due to time constraints. (J) is straightforward as written, but matters of access may become more complex than this implies in real life.

Illustrative Case 1—Description. The evaluand is a career education program being piloted nationwide. Three schools districts in each state were to participate in the field test. Districts submitted proposals that were rated by four judges on 13 separate criteria. The criteria themselves are not mentioned, so it is impossible to tell if or how cultural diversity was considered in the selection process. The mean ratings were used to rank order the proposals, and the top three were then selected. The author of a proposal that was not awarded a contract pursued the matter by traveling to Washington, DC, speaking to a government official and demanding an audit of the ratings. The official complied, and the audit revealed that in this and five other cases, the mean rating had been incorrectly calculated, resulting in erroneously low rankings for six proposals.

Illustrative Case 1—Analysis. Not surprisingly, the analysis focuses on the failure to check the accuracy of the calculations, suggesting that two staff members calculate the means independently. This is a mechanical failure, one that does not immediately signal cultural relevance. It does feel dated to me, however. I picture staff with hand calculators entering columns of figures rather than an Excel spreadsheet.

Illustrative Case 2—Description. This case has potentially greater social relevance, though it is written up to illustrate a coding error detected via external audit. The evaluand is child support guidelines enacted by state legislation. The evaluator is a law professor from a different state. The evaluation sought to determine whether lawyers, judges, and family mediators were implementing the guidelines similarly and whether the guidelines were affecting other terms of settlement in disparate ways. While this latter question raises potentially interesting equity issues, the case describes only data pertaining to the first question. Data were collected via interview, based around three hypothetical cases. Interviews were audiotaped, transcribed, and reviewed/corrected by the interviewers who were law students in the state enacting the legislation. The evaluator reviewed the transcripts, analyzed the data, and wrote a report concluding that the judges were not consistently implementing the mandated guidelines. The state bar association demanded an audit. While overall consistency among transcripts, findings, and conclusions was verified, the auditor discovered that one interviewer had reversed the respondent codes for judges and mediators. It was the family mediators, not the judges, who were inconsistent in their implementation. The evaluator published a retraction of the original findings and apologized to the judges.

Illustrative Case 2—Analysis. The analysis zeros in on the coding problem and how it could have been prevented by (a) more carefully training the interviewers to check their transcription codes; (b) checking transcriptions against tapes to assure proper attribution; (c) hiring a non-interviewer to listen to each interview while reading the transcript, checking all assigned codes; and/or (d) the evaluator himself checking for transcript errors before analyzing data. The analysis does not applaud the evaluator's training of the interviewers that produced well-documented, professional interviews, clear responses, and quality transcripts nor does it point out that it was the careful preservation of the original tapes in sealed cartons that permitted the audit to occur. Once the error was discovered, the evaluator acted appropriately to retract and correct his original findings. In hindsight, the error could have been prevented, but the analysis itself could have been more balanced, as there was much for which the evaluator should be commended. Cultural dimensions of information control are not addressed in the description or the analysis.

Supporting Documentation.

Crevecoeur, D., Finnerty, B., & Rawson, R. (2002). Los Angeles County Evaluation System (LACES): Bringing accountability to alcohol and drug abuse treatment through a collaboration between providers, payers, and researchers. *Journal of Drug Issues, 32*, 865-879.

Return to Table of Contents

A8 Analysis of Quantitative Information.

Quantitative information in an evaluation should be appropriately and systematically analyzed so that evaluation questions are effectively answered.

Standard. Assuming that quantitative information is relevant to answering the questions that have been posed, the standard itself seems appropriately written.

Overview. Age and socioeconomic characteristics are given as examples of potentially useful quantitative data concerning respondents. Process and outcome measures are listed generically, without examples, though clearly some of these data could be culturally relevant as well. The overview advocates exploratory analysis, followed by "more sophisticated and complex analyses" to provide clear results. The overview encourages the use of visual displays. While these are necessarily vague, general statements, it would seem appropriate to distinguish between use of descriptive versus inferential statistics and to steer clear of advocating "fishing expeditions" such as those looking for race or gender differences in the absence of any theoretical justification for doing so.

The third paragraph addresses the need to examine differences between non-randomly assigned comparison groups, cautioning against using statistical corrections for initial differences without scrutinizing underlying assumptions. The focus is on independent variables, and the presence of intervening variables such as age or socioeconomic status, race or gender is not addressed. The standard would be strengthened by adding content on the appropriate treatment of such data in quantitative analysis. The final overview paragraph opens the door for such discussion by introducing the possibility of subgroup differences and the need to examine variability as well as mean effects. The caution that evaluators must be able to defend their choice of method, including statistical calculations and their underlying assumptions, is congruent with Davis' (1992) argument concerning the use of race as an explanatory variable. Evaluators should be advised to consider variability within demographic subgroups as well as differences among them and to report similarities across groups rather than only highlighting differences.

◙ Acknowledge cultural context explicitly in revising this standard.

Guidelines. (A) I would add, ". . . and the cultural diversity of the context being studied." This may be reflected in "the nature of the data," but I think it is important that the evaluator be reminded to notice relevant cultural variables in planning and conducting the analyses. (B) is standard practice and certainly appropriate. Here, or as a separate Guideline, might be an appropriate place to advise evaluators to explore diversity within cultural subgroups and similarities among minority and majority groups rather than only analyzing difference between cultural sub-groups. (C) is important and should explicitly include weaknesses that limit the full representation of diverse populations (e.g., groups that were dropped from the analyses due to a small number of respondents). (D) Effect sizes merit more discussion than appears here, reflecting advances in statistical analysis and changes in the way in which statistical significance is understood. I like the recognition of practical significance, but to me this extends considerably beyond effect sizes to include real-world impact on lived experience. This is a key element of both consequential and experiential justifications of multicultural validity. I would expand the illustrations of practical significance here.

Common Errors. (B) echoes the need to separate considerations of practical significance from those of statistical significance. I agree with this, but then disagree with (C) which undermines consideration of practical significance. Replicability is not always the predominant criterion, depending on the nature and purpose of the evaluation. It may, in fact, bias evaluation of programs tailored to distinct cultural contexts, as illustrated by meta-evaluation of PUSH/Excel (House, 1988; Stake, 1986). (D), (E), and (F) are procedural errors of calculation, though (F) may be particularly relevant to the proper application of subgroup norms and the use of standardized educational tests in "high stakes" situations. (G) and (H) raise interesting contrasts between complexity and simplicity, rigor and relevance, though I think that one has to be careful not to create false dichotomies and particularly not to imply that certain stakeholder audiences are necessarily incapable of understanding the results of complex analyses *if they are clearly explained*. Definitions of "rigor" need to be recast to include cultural relevance. (A), (I), and (J) are errors of conceptualization more than analysis, but they are certainly errors worth citing. (K) should be updated to reflect advances in mixed-method evaluation

(Greene & Caracelli, 1997). ◙ Add the Common Error, "Allowing the quantitative analysis of the program to be distorted because qualitative factors like culture are inaccurately captured ." ◙ Another potential Common Error, "Failing to disaggregate the data for subgroup analysis, which would capture differences in subgroup behaviors and performance."

Illustrative Case 1—Description. The evaluand is a five-week project to improve reading performance of low-achieving fourth grade children. The evaluators are an external company. Children were screened into the program based upon performance on a standardized reading test. Posttests were administered to participants and non-participants, and gain scores were calculated and average gain scores compared between groups. The composition of the comparison group is not specified, though it presumably included fourth graders who had scored above the cutoff on the pretest as they are referred to as "better readers". No cultural context information is given.

Illustrative Case 1—Analysis. The analysis points to statistical regression as a threat to the validity of the evaluators' conclusion that the special project improved the reading scores of poor readers. This is more an error of design than of analysis; the recommendation is that low scoring students should have been randomly assigned to two groups, one of which received the five-week special project and one that did not. (Additional threats to validity such as diffusion of intervention or compensation of the control group are not addressed.) The analyst recommends three separate analyses be compared—posttest data; gain scores, and ANCOVA—and that graphic techniques be used enhance data interpretation. There is no discussion of the practical significance of the project and no recommendation that this be considered.

Illustrative Case 2—Description. Gender is explicitly addressed in the second case. The evaluand is a high school program intended "to encourage girls to enroll, participate, and achieve in physical science courses." This case also provides context information. The program is delivered in different formats across schools in the district. Physical science courses in some schools are co-educational; other schools offer special sections for girls; still others offer girls the choice of co-educational or gender specific sections. The illustrations given of variables in the study include parental education and gender of the instructor, dimensions of diversity that appear relevant to this particular program. The district contracted with an external evaluator for the statistical analysis, and the contractor used hierarchical statistical regression to compare girls-only and coeducational science instruction, with and without choice. Results were reported a fractions of standard deviations. Same-gender programs outperformed the coed program one fourth of a standard deviation on the physical science tests and one-fifth of a standards deviation on the criterion, "enrollment in a science-related course in the following year." Parent education accounted for 15% of the variance among class achievement means. The contractor suggested that increasing parental education by one standards deviation would improve class mean achievement and decrease the gap between same-sex and coed courses.

Illustrative Case 2—Analysis. The statistician didn't appear to focus on the evaluation question of primary interest, and the suggestion regarding increasing parental education seems a bizarre departure from the policy decision at hand. The analyst is understated in noting, "It is possible that the analysis did not address issues within the control of policy-makers." The analyst states that the data were analyzed correctly (a fact that is difficult to

determine from the information given). In fact, the design itself appeared to be a two by two factorial (co-ed/non-coed x choice/no choice) in the description, three comparison levels (“single sex, coeducational, and choice”) in the analysis. The analyst asserts that these three kinds of program are confounded with various combinations of background and instructional characteristics, presenting an inaccurate picture of how a school would operate under “unconfounded circumstances.” I don’t take the point here. In terms of ecological validity, where exactly would one find “unconfounded circumstances” operating in the public schools? The point about policy decisions and informed debate is an interesting one, however. If one accepts the goal of statistical analysis as informing debate, as the analysis argues, then the choice of statistic may be constrained by the ability of policy makers to understand data. Yet it seems to me that to include diversity dimensions appropriately often entails the kinds of complex multivariate analyses that the analyst here recommends against in the interest of facilitating debate. To restrict the analyses and omit relevant considerations of cultural context would also curtail debate, perhaps in ways not obvious to policymakers.
◙ An Illustrative Case could describe analyses that failed to account for subgroup differences.

Supporting Documentation.

Davis, J. E. (1992). Reconsidering the use of race as an explanatory variable in program evaluation. In A. Madison (Ed.), *Minority issues in program evaluation*. *New Directions for Program Evaluation*. No. 53 (pp. 55-67). San Francisco: Jossey-Bass.

Greene, J. C., & Caracelli, V. J. (Eds.) (1997). *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms. New Directions for Evaluation*, No. 74. San Francisco: Jossey-Bass.

House, E. R. (1988). *Jesse Jackson and the politics of charisma: The rise and fall of the PUSH/Excel program*. Boulder, CO: Westview Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton-Mifflin.

Stake, R. E. (1986). *Quieting reform: Social science and social action in an urban youth reform*. Champaign, IL: University of Illinois Press.

Weisburd, D., Lum, C., & Yang, S. (2003). When can we conclude that treatments or programs ‘don’t work’? *The Annals of the American Academy of Political and Social Science, 587,* 31-48.

Return to Table of Contents

A9 Analysis of Qualitative Information.

Qualitative information in an evaluation should be appropriately and systematically analyzed so that evaluation questions are effectively answered.

Standard. As in A8, one must assume that the information was well selected and is relevant to the evaluation questions posed. If this is the case, then the standard itself is appropriately stated. This standard is especially relevant to cultural diversity, as many culturally congruent methods use qualitative information.

Overview. The first paragraph defines qualitative data and describes parameters of how it may be collected and recorded and on what program dimensions it may focus. It also makes the interesting point that it may be gathered intentionally or unexpectedly, one of the few references the Standards makes to emergent designs. This is further elaborated in the third paragraph, which describes the inductive, iterative, and interactive nature of qualitative analysis, including intuitive analysis that extends beyond pre-specified rules. The overview makes clear that procedures exist for confirming and verifying qualitative data. In addition to seeking confirming or disconfirming evidence from more than one source and subjecting inferences to independent verification, iteratively re-checking with the same sources to make sure that they were correctly understood and represented is a typical verification procedure. The description of appropriate qualitative analysis (p. 172) is limited to the construction, verification, and interpretation of categories sufficient to answer evaluation questions. While this is certainly one type of qualitative analysis, it is by no means the only type. In the Humanities, for example, one can apply critical theory or queer reading to text material in a way that is not as classificatory as the procedure described in the overview. Just as the quantitative analysis standard alludes to many types and levels of analysis, A9 should be written to communicate that there is more than one appropriate strategy for analysis of qualitative information.

I was also struck by the closing statement of the overview indicating that this standard is intended to safeguard evaluation from inappropriate analysis that may lead to premature closure and inappropriate crosschecking. While the same could be said of quantitative analysis in A8, it was not. Apparently qualitative information is seen as riskier or more dangerous; evaluators need to be armed with safeguards when approaching it. The subtle privileging of quantitative over qualitative information is something that I think we should challenge.

Guidelines. (A) As in A8, I would add, “and to the cultural context of the study.” (B) Not all methodologies consider single-source data as a weakness. For example, if one is doing life histories or seeking to give voice to a particular person’s story. There may also be “irreconcilable” contradictions across or within cultural subgroups that should be recognized rather than erased; I would not consider this a weakness either. (C) sets a preordinate design for qualitative work, but this is not the only option. Not all qualitative methods would define the parameters listed up front. (D) equates analysis with categorization, as was done in the overview. While this is one analytical strategy, it is not the only analytic choice for qualitative data, and I would hate to see categorization equated with validity (A5 Valid Interpretations) as is implied here. (E) speaks to reliability checks that are appropriate for categorical analyses, but inter-rater reliability and external audit are not the only possibilities for assuring consistent, trustworthy interpretation. (F) is a potentially very important guideline that could encompass primary inclusion of direct and indirect consumers and diverse stakeholder groups. It’s interesting that this is presented as a credibility issue (U2 Evaluator Credibility). I see it as a central validity issue (A5 Valid Interpretations), elevating its importance. (G) Triangulation is generally a good principle, though its role and operationalization may vary in different types of inquiry. For some types of data (qualitative or quantitative, there may be a single best source). (H) is important and should not be buried at the end of the list!

When the standards are rewritten, careful attention should be paid to epistemological diversity of the writing/review panels as well as cultural diversity. The two often intersect but are certainly not equivalent. Nevertheless, standards that are broadly constructed to be inclusive of multiple epistemologies will support multicultural validity by affirming that responsible, high quality evaluation may be done outside of a majority perspective.

◙ Add the Guideline, "Acknowledge evaluators are not fully objective, and prepare them for qualitative analysis with proper training in technical skills and the influence of contextual factors."

Common Errors. (A) is a run-on sentence that appears to be two separate Common Errors. It should be divided at the comma. I strongly agree with the first idea (A1) regarding the error of seeing qualitative data as non-rigorous. Regarding (A2), *no* research is free from preconceptions, and the validity of those theories, preconceptions, working hypotheses should always be questioned, be it in qualitative or quantitative research. All designs should provide opportunity for disconfirming information. **This is another example that makes me question whether the qualitative/quantitative dichotomy is still the most relevant structure for organizing analytic and interpretive standards.** It's true that mechanical and procedural differences exist, but the A8/A9 division also seems to perpetuate some false dichotomies at the level of principles. What do the rest of you think? **(B) is extremely important to cultural competence.** The lived experiences of persons from cultural majority/minority groups may be very different and hence their values and world views may also differ. "Reality" should not be taken as a singular fixed perspective. (C) is also extremely relevant. Cultural competence should be included in considerations of "degree of expertise." (D) I certainly agree that qualitative and quantitative data are complementary and that they hold in common many principles of analysis. My only concern is that the last phrase of (D) may be read as a call to always accompany qualitative data with quantitative (or vice versa) which may be too restrictive. While no one would argue against maintaining a balanced perspective, (E), (F), and (G) seem to be worded in ways that could be used to argue against attention to diversity. I am particularly concerned that relevance and rigor are presented as opposing forces in (G). Cultural relevance should be considered a necessary component of rigor. (H) Certainly major "redirection" would require discussion, but the contract may well permit exploration of questions emerging from the data without explicit re-review. The contract should clarify the nature of oversight that the client wishes to have of the data analysis. (I) The volume of data collected—qualitative or quantitative—should be governed by the resources available for the evaluation. (J) and (K) speak to concerns I raised earlier. Not all qualitative data analysis involves categorization and quantification. These are important points that should not be buried at the end of the list of Common Errors.

◙ Add the Common Error, "Failing to recognize the direct influence of the evaluator's perspective in assessing the program."

Illustrative Case 1—Description. The evaluand is a crime prevention project. The evaluators are two members of a school district's evaluation office. No information is given on the cultural context of neither the program nor the personal characteristics and expertise of the evaluators (which apparently did *not* include skill in qualitative analysis according to the scenario). The evaluators were given a wealth of qualitative information from staff logs, archival records from school and corrections, testimony taken at quarterly hearings, media coverage, and minutes of community advisory board meetings. The superintendent wanted to be able to accurately represent the evolving nature of the project's activities, and the funding agency wanted to know if contacts between the youth and various law enforcement agencies had been reduced. The evaluators chose to develop inclusive categories under which to organize all of the qualitative data and to do simple frequency counts of contacts with law enforcement. The categorized data were of minimal help in documenting the changing nature of project activities over the year, and the frequency counts revealed no difference in number of contacts from the year prior to the program. Based upon these findings, the funders cancelled the program. Project staff pointed out that the nature of the contacts with law enforcement had changed from negative (arrests and trials) to positive (counseling and supervision) and that the corrections staff viewed the project as successful.

Illustrative Case 1—Analysis. The analysis focuses on the methodological incompetence illustrated in this case, noteably the use of simple frequency counts to communicate contact with law enforcement. The culturally bound assumption that contacts with the law would be negative—at least for these youth—was not remarked upon. The analyst emphasizes the need to focus the evaluation on questions of interest and prioritize data analysis rather than trying to analyze all available data. The analyst does not question how the values, background, training of the evaluators may have shaped their quantitative focus. Clearly these evaluators did not attempt to capture the worldview of the youth who were participants in this evaluand nor their view of the problem that the evaluand was designed to address. The "detailed case study" data were not examined. If more cultural information were available, it would permit deeper reflection on the dynamics of position in this illustrative case. Absent such, one is left with the impression that these evaluators were inexperienced and inept at handling qualitative data, surprising given their district-level position.

Illustrative Case 2—Description. The context is an American law school; the evaluation is funded research by a law professor seeking to determine the adequacy of case method in teaching law. The evaluand is actually "the methods used by federal judges in pretrial settlement conferences." Since this evaluand is not a program, it is a weak example for use in these particular standards despite the illustrative use of qualitative data. The data consisted of observations and recordings of pretrial conferences, and interviews (also recorded) with federal district court judges. The evaluator (presumably the law school professor) was skilled in quantitative methods and was presumably unfamiliar with qualitative methods. No cultural descriptors are provided. He/she initially consulted a qualitative methodologist but found that consultant's data reduction and analysis procedures too time consuming. Instead, the evaluator chose to transcribe and edit the tapes for flow and to analyze the data using "an iterative, intuitive process." The results confirmed what is implied to have been the evaluator's working hypothesis—that case method and current pedagogy are inadequate preparation for modern legal practice.

Illustrative Case 2—Analysis. The analysis carefully enumerates four guidelines that were violated in this case. [Note: though this illustrative case is new to the 2nd edition, the author must have been referring to the 1st edition of the *Standards* in writing, "The evaluator in this case violated four of the six guidelines for the handling and analysis of qualitative information." There are *eight* A9 guidelines in the 2nd edition, not six.] Violations included: tampering with the raw data by editing transcripts to improve flow;

employing nonsystematic analysis procedures, based upon intuition; failing to budget sufficient time/resources for an appropriately systematic analysis; and failure to consider alternate interpretations or disconfirming information. The case illustrates how the values of the evaluator may create a self-fulfilling prophecy when appropriate analysis procedures are not followed. Because only pedagogical background is provided, this is all that is visible in the analysis; however, were cultural background available, it might shed light on additional value-based assumptions that permeate this research.

Supporting Documentation. ◙ Denzin & Lincoln (2nd ed.) should be cited here, or appropriate chapters therein.

Gliner, J., & Sample, P. (1996). A multimethod approach to evaluate transition into community life. *Evaluation and Program Planning, 19*, 225-233.

Johnsen, J., Biegel, D., & Shafran, R. (2000). Concept mapping in mental health: Uses and adaptations. *Evaluation and Program Planning, 23,* 67-75.

Return to Table of Contents

A10 Justified Conclusions.

The conclusions reached in an evaluation should be explicitly justified, so that stakeholders can assess them.

Standard. This is a clear statement of the importance of justifying claims and of stakeholder inclusion in assessing such justifications, but it is vague regarding *what* stakeholders are assessing, which is, ultimately, validity.

Overview. I like the inclusive definition of conclusions, covering both judgments and recommendations. This parallels Messick's unified validity theory, which gives attention to both inferences and actions. These conclusions must be both defensible and defended—the logic must be explicated, not implied or assumed. Access to such explanation is key to understanding the culturally bound assumptions that underlie conclusions, making this a particularly relevant standard from the perspective of culture. Are the judgments and actions flowing from an evaluation warranted? Despite the reference to "underlying assumptions," I find the attention to values to be insufficient. This standard is fundamentally about data synthesis: interpretations of data through values to reach conclusions. To me the values dimension is not clearly communicated. The "plausible alternative explanations" may also reflect cultural standpoints; therefore, it is particularly important to articulate why these were rejected. The restrictive clause, "where possible" in this last sentence seems odd. To me, it is *always* possible to reflect on alternate interpretations, and it is important to do so (A12 Meta-evaluation).

Guidelines. (A) The notion of "faithful reflection" frames this as a validity issue, and I strongly agree with this representation. I also like the link to both questions and to procedures and data. The conclusions should be justified in both frames of reference. (B) again explicitly makes this a matter of validity. Well put! In (C), the plausible alternative explanations should certainly be considered, but the rush to "discount" them may proliferate bias. This guideline seems to encourage discounting, setting it as a desirable goal “where possible.” This may not be in the best interest of thoughtful reflection, especially where diverse cultural perspectives need to be considered. (D) cautions against overgeneralizing. I agree that the limits of generalizability should be addressed, including generalizability along majority/minority population dimensions or across culturally defined dimensions of diversity. (E) Caution is always advisable. Whether or not the particular findings are considered equivocal, they are in the Big Picture incomplete, partial, or temporary understandings, pending further study. Cronbach affirmed this in his discussion of validation as an open-ended process, and we would do well to remind evaluators of that in this context also. Evaluation, like validation, is open-ended. (F) underscores the importance of tapping diverse perspectives, including direct consumers/participants, emphasizing formative feedback prior to finalizing the report. “Common misinterpretations and inappropriate inferences” may include the (mis)use of race as an explanatory variable. This guideline does not address the fact that “credibility” is itself a culturally-bound construct.

Common Errors. (A) refers to limitations of procedures and of data, but limitations of *perspective* should also be explicitly cited. Such fundamental limitations may color the selection of both methods and procedures, but they may also infuse the evaluation more subtly with unacknowledged bias. (B) I agree strongly with the importance of capturing unintended outcomes (Scriven’s preferred term for “side effects”). (C) sounds straightforward—conclusions not grounded in sound, sufficient information constitutes an error, but the key is how much is sufficient? Often “sufficient” is viewed as the minimum information necessary for decision-making. Different definitions/ decision rules apply for what is deemed sufficient—e.g., between managers and academicians. The same applies to definitions of “soundness” (validity). As I’ve argued elsewhere, justifications of multicultural validity may be methodological, interpersonal, consequential, or experiential (and there are no doubt other relevant justifications yet to be explored). The judged soundness of conclusions may differ depending upon the justification invoked. (D) To me, the admonishment against being too cautious relates to the danger of evaluation having no impact. I would cite U7 Evaluation Impact as supporting this concern. (E) Both strengths and limitations should be cited. This is fundamental to meta-evaluation, and A12 should be referenced. Unrecognized limitations may include noting infusions of power and privilege, including the influence of funders and sponsoring organizations.

Illustrative Case —Description. The evaluand is a K-12 program in ecological education. The evaluation is federally funded, and the contract was awarded to “an evaluation firm” of unspecified credentials or disciplinary background. The test sites were geographically diverse: three urban, four suburban, and ten rural districts. No other information on cultural context of the programs or of the evaluators is provided. The federal funding agency was considering widespread dissemination of the program. They wanted to know if the program was meeting its (unspecified) objectives, if teachers were incorporating ecology ideas in their existing courses, and how curriculum specialists rate the program compared to other ecological education models. The funders also directed the evaluators to examine the results of pilot study completed while the program was under development.

The evaluation design is described as consisting of four parts: a test-retest design completed annually using the program’s end-of-course test; “periodic discussions” with groups of students, teachers, and administrators concerning program implementation; annual interviews with teachers and curriculum specialists in each participating school; and a cost analysis. Results reported focused on the pretest-posttest gains, which were significant across all grades and schools for both years. The evaluators recommended dissemination. Funders rejected the report, arguing that it omitted attention to results of other design elements (and presumably to some of their evaluation questions, though the tenuous link between funders' questions and the evaluation design elements was never explored). Presumably these data had been included in an Appendix to the report, as had the pilot study.

Illustrative Case—Analysis. The brief analysis notes the obvious: evaluators should have considered all available data in formulating their recommendations and conclusions. The reviewer notes that certain elements—notably the cost analysis—apparently went unreported. Interestingly, cost was not one of the evaluation questions reportedly posed by the funder, so the impetus for its inclusion in the design is unclear (which may relate to its not being reported). The alleged relevance of the pilot study remains questionable to me, however. The funders clearly wanted it included, and the evaluators complied by including it in the Appendix to their report. The evaluators had questioned the utility of those pilot data, however, because they had been collected while the program was under development and did not necessarily reflect final implementation. Minimally, the (unspecified) sampling frame of the pilot study would have been non-comparable. I would agree with their logic and question the motives (values, underlying assumptions, logic—all elements relevant to this standard) that led the funders to "push" this material.

Supporting Documentation. ◙ The supporting documentation listed is particularly weak. The Hendricks & Pappagiannis reference speaks to the mechanics of presenting recommendations much more than to scrutinizing their underlying logic and assumptions. The Smith reference seems tangential.

Carroll, J., & Doherty, W. (2003). Evaluating the effectiveness of premarital prevention programs: A meta-analytic review of outcome research. *Family Relations, 52*(2), 105-118.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.) *Measuring achievement: Progress over a decade. New Directions for Testing and Measurement*, *No. 5*, 99-108. San Francisco: Jossey-Bass.

Davis, J. E. (1992). Reconsidering the use of race as an explanatory variable in program evaluation. In A. Madison (Ed.), *Minority issues in program evaluation, New Directions for Program Evaluation,* No. 53, (pp. 55-67). San Francisco: Jossey-Bass.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Naylor, P., Wharf-Higgins, J., Blair, L., Green, L., & O'Connor, B. (2002). Evaluating the participatory process in a community-based heart health project. *Social Science and Medicine, 55*, 1173-1187.

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

Stratton, K., & Delaney, J. (2000). Reviewing changes to the child disability allowance: Giving parents a voice. *Australian Social Work, 53*(2), 5-11.

Return to Table of Contents

A11 Impartial Reporting.

Reporting procedures should guard against distortion caused by personal feelings and biases of any party to the evaluation, so that evaluation reports fairly reflect the evaluation findings.

Standard. The current presentation of this standard frames impartiality as a micro issue—the personal feelings, values and biases of individuals. In so doing, it understates the range of distorting influences, which may include macro (e.g., systemic heterosexist bias) or mezzo (e.g., institutional racism) issues versus individual homophobia or racism felt or expressed by individuals. It may also include individual distortions of which the person is unaware (e.g., white privilege). This standard is extremely important to issues of justice and fairness that intersect culture, but it is quite a bit more complex than its current wording.

Overview. The overview appropriately frames distortion as a limitation of perspectives, yet still implies that it results from a personal flaw—carelessness or inability to resist pressure—rather than systemic bias that may infuse a report (e.g., societal attitudes toward age or disability). Given the complexity of these issues, the overview seems skimpy and underwritten. It also seems to single out formative evaluation procedures of continuous reporting and ongoing program improvement, implying that formative evaluation is somehow more subject to distortion than is summative. The point that the evaluation process itself may make it more (or less) subject to certain kinds of distortion should be broadly stated and illustrated rather than singling out a single model for critique.

Guidelines. (A) I agree the fair reporting should be addressed with the client initially, but such understandings will likely need to be reaffirmed and/or operationalized as the evaluation unfolds. As written, (A) implies that this is something that can be "handled" at the outset and then set aside. This is inconsistent with the spirit of the standard. (B) "Editing" should be broadly defined to include potentially interactive reviews/comments/suggestions by multiple stakeholders. I agree that final authority/responsibility for content should be clear. Such authority should also be considered within its cultural context. **Note that all of the standards that address reporting, including this one, seem biased toward *written* reports, not always the most culturally appropriate choice.** (C) overstates the "independence" of perspectives along the internal/external continuum. *Degrees* of connection, investment, and proximity to or distance from the evaluand should be noted rather than portraying it as a dichotomous independent/dependent distinction. Since the inclusion of multiple perspectives supports validity; I would also cite U3 Information Scope and Selection, and A5 Valid Information here. (D) Point well taken. It gives important attention to alternative interpretations and recommendations with appropriate cross-listing of A10 Justified Conclusions. (E) Strategies to maximize the recognition of contrasting perspectives are noteworthy, though not all are equally appropriate to a given model or context. For example, rotating team members could compromise trust and rapport or

introduce instrumentation threats to internal validity, depending upon one's design. (F) also relates to Systematic Information Control (A7) and Meta-evaluation (A12).
Common Errors. (A) is very important. Assumptions of "neutrality" should always be interrogated. (I think it's an error to assume that *any* parties to an evaluation are neutral.) (B) is an error in the broadest sense, but here again, distortion is cast as personal failure, implying that all distortion can be eliminated via "safeguards." Systematic Information Control (A7) can support the intent of this standard, but "distortions" may ultimately lay in the eye of the beholder and his/her majority/minority position. (C) This is an interesting error—one that presumes that authority *should* remain exclusively with the evaluator and that any shared responsibility is a violation of standards. This position is inconsistent with certain collaborative/participatory/empowerment models of evaluation. The standards should not privilege certain models over others. (D) Taken out of context, this seems somewhat random, although public disclosure and communication with right-to-know audiences should certainly be provided for. (E) Certainly keeping lines of communication open in ways consistent with the operational model is important. But again, this "error" may not apply to all models—for example, goal-free evaluation. (F) I agree that both the client and the evaluator must consider the possibility of negative findings and alternative interpretations thereof.
Illustrative Case 1—Description. The evaluand is a special reading project, elementary level, piloted in one elementary school for three years. We are told that this is a small district, but no other context information is given. The initiator of the evaluation is the district superintendent and the evaluator is a reading specialist from a neighboring district. No other information is given regarding this woman's credentials or personal characteristics, but she allegedly preferred highly structured approaches to teaching reading and formed premature judgments ("as she *began* collecting data" [emphasis added]) that the program—which was *not* highly structured—was unsuccessful. She presented her impressions to staff and found them to be "very defensive and hostile." The staff asserted that heavy reliance on test scores had constrained accurate understanding, and they questioned the evaluator's predisposition toward structured reading programs (U2 Evaluator Credibility). The evaluator changed her report and recommended adoption of the program, a shift that was described as "capitulating to pressure."

Illustrative Case 1—Analysis. The analysis faults the evaluator for her inexperience, a point not mentioned in the case description. The analysis points to overreaction on the part of the evaluator in "capitulating" to the views of the program personnel, but rather than representing an impartial stance as called for by this standard, it seems to be heavily *biased* against the program perspective. Program staff, we are told, "had an axe to grind," though nothing in the description suggests this. The analyst recommends that the program persons' reactions be discounted because they were committed to their program—hardly a respectful stance. The analyst takes an arrogant, authoritarian perspective, elevating the views of the evaluator above those of the program personnel. Triangulating data sources to support validity is necessary and appropriate, but that can be accomplished without impugning the credibility of internal perspectives. The value positions (and assumptions of impartiality) of the evaluator should be examined with the same degree of scrutiny given the program personnel and other stakeholder perspectives. **Both the case and its analysis need revision.**

Illustrative Case 2—Description. The evaluand is a library of videotapes and support materials compiled by a large, state medical school. No other context information is given. The materials document emergency room cases and are to be used to train medical students in trauma care. A   faculty committee created the library over a two-year period, assisted by a panel of outside experts hired to evaluate the development process and materials produced, both formatively and summatively. When the panel of experts came together to compile their respective sections of the final report, the chair challenged the impartiality of one member and asked him to remove his content from the final report. She apparently persuaded a majority of the fellow panelists to support this decision, though not all agreed with her stance. We are told only that is "apparent" to the Chair that the panelist "had taken his 'formative feedback role' too seriously" and had used his expertise in producing nonprofit documentaries to advise the library committee. She judged this to be "too involved in the project to be an objective evaluator."

Illustrative Case 2—Analysis. The analysis applauds the action of the panel chair in editing out a portion of the final report when she judged the impartiality of the panel member to have been compromised. No mention is made of the gap this left in the report or the omission of the perspective of expertise for which this panelist had been hired. No mention is made of the chair's failure to interrogate her own objectivity or that of the other panelists. No notice is taken of the pressure exerted on panelists to support her view of what was "apparent." Nor does the analysis address the reaction of the medical school faculty committee who had hired the panelists. At best, the case gives incomplete justification for a dubious action, applauded by the analyst who sees overinvolvement as a common "trap" of formative evaluation—another clear instance of bias against participatory/collaborative/empowerment models. Without knowing more background information, including cultural context of both individuals and organization, it is impossible to know what other political or personal factors may be in play in silencing the voice of the panelist.

Like the first case, this second illustrative case does not strike an impartial stance; it is biased against formative evaluation in general and participatory models in particular, and it should be revised.

Supporting Documentation.

Booysen, F., & Arntz, T. (2003). The methodology of HIV/AIDS impact studies: A review of current practices. *Social Science and Medicine, 56*, 2391-2405.

Kovacs, P. (2000). Participatory action research and hospice: A good fit. *The Hospice Journal, 15*(3), 55-62.

Wholey, J. (1997). Clarifying goals, reporting results. *New Directions for Evaluation ,* No.76, (pp. 95-105). San Francisco: Jossey-Bass.

Return to Table of Contents

A12 Metaevaluation.

The evaluation itself should be formatively and summatively evaluated against these and other pertinent standards, so that its conduct is appropriately guided and, on completion, stakeholders can closely examine its strengths and weaknesses.

Standard. This is an extremely important standard. Cultural critique should become a routine part of meta-evaluation. Here, the treatment of formative and summative is even-handed; I like the dual emphasis. I would say, "conceptualization, implementation, and interpretation are appropriately guided…" to emphasize the breadth of scrutiny that is desirable. Since strengths and weaknesses may be closely examined both formatively and summatively, I'm not sure that the attention drawn to "on completion" is necessary.

Overview. The overview of this standard is well developed. The first paragraph lays a rationale for evaluating evaluation, pointing to evaluation's potential role as a change agent or in support of existing programs. I agree that evaluation is difficult to do well (particularly with regard to multicultural validity) and that it offers the potential to change systems and improve our ability to work for social justice. I would like to see the consumer brought into the conversation. Citing P1 Service Orientation, I would emphasize evaluation's potential to make systems more responsive to consumer need. As written, the emphasis on decision makers gives it a management bias.

The second paragraph gives the definition of meta-evaluation and discusses the role of both internal and external meta-evaluation. I think it's important to recognize that there is a continuum of proximity to/distance from the evaluation and that meta-evaluation can be performed by persons at different locations on that continuum. For example, the client of the evaluation—who, I would argue, implicitly engages in a type of meta-evaluation whenever he/she reviews a project for contractual compliance—may be external to the implementation of an evaluation but internal to its design and conceptualization by virtue of having providing the framing questions and signed off on their operationalization. The last two sentences of that paragraph seem a little out of context insofar as each is speaking to a specific benefit of meta-evaluation. I would move them to follow the elaboration of formative and summative meta-evaluation to avoid disrupting the flow of that argument. Benefits of meta-evaluation should be expanded to include supporting cultural competence and enhancing multicultural validity.

Paragraph three expands on formative meta-evaluation, and here it errs in understating the value of internal formative meta-evaluation. To say that formative meta-evaluation is "ideally" external undermines a respected literature on reflective practice (cf., for example, Schön's work). This is particularly troubling since all theories of cultural competence with which I'm familiar include a reflective component. It's key to culturally competent evaluation that evaluators be told that standards of good practice requires evaluating their work while they are doing it. Internal meta-evaluation should *not* be framed as the default, to be used only when resources are too scarce to mount an external meta-evaluator. **I take strong exception to the message of this third paragraph.**

Paragraph four defines summative meta-evaluation, and here the treatment of internal and external roles is more balanced. I like the explicit mention of multiple perspectives, though the idea could fruitfully be extended to internal as well as external perspectives, and the introduction of formal versus informal meta-evaluation. Though the terminology risks setting up a false dichotomy, I do think it's important to point out that there are many different levels of depth or intensity of meta-evaluation just as there are in evaluation itself. Interestingly, the example given of informal meta-evaluation using the standards checklist is an internal, formative example, a contradiction of the message of the previous paragraph. Because the fourth paragraph also begins introducing questions that could be posed in and answered by meta-evaluation, it would be appropriate to include questions about the cultural congruence of the evaluation design, the respectful inclusion of stakeholder audiences, the degree to which multiple value frames were considered, etc.

Paragraph five enumerates audiences for meta-evaluation with appropriate breadth, though the sequence of presentation could be read as privileging the decision makers. It might be clearer to pair the audience discussion with the discussion of benefits of meta-evaluation.

Paragraph six enumerates many benefits of meta-evaluation (so the benefits in paragraph two would fit nicely here), but supporting culturally competent, multiculturally valid evaluation is not on the list. This should be explicitly addressed under the umbrella of advancing state-of-the art evaluation. A multicultural society demands no less. Because the last sentence speaks to a benefit of formative meta-evaluation only (tying it to evaluability assessment), I would move that up to paragraph three.

Guidelines. Though (A) gives equal attention to formative and summative evaluation, the Guidelines, taken as a whole, apply more to summative than to formative and appear to have been written from a summative perspective. (B) Assigning responsibility for the function is not a bad thing, but I think it's important to frame meta-evaluation—especially formative meta-evaluation—as *everyone's* responsibility. For example, each member of the evaluation team or stakeholder advisory board has a unique vantage point on cultural context by virtue of his/her personal characteristics, values, and lived experience. Conversations about potential cultural bias or limitations of perspective should be encouraged as a routine component of project management, alongside conversations about timelines, response rates and budget. (C) scratches the surface of a major issue. As in any evaluation, the credibility of the meta-evaluator is enormously important. And as noted in the critique of U2 Evaluator Credibility, such judgments are saturated with values that must be understood in cultural context. By focusing on procedures through which the chair of an external meta-evaluation team might be selected, (C) seriously constrains, and in my opinion trivializes, considerations of credibility. Cultural competence of the meta-evaluator (or meta-evaluation team) should be included under credibility. (D) applies more to summative evaluation than to formative and to formal meta-evaluation rather than informal. (E) is an oddly inappropriate procedural recommendation that goes beyond the guidance of (D) to state a very specific rule. While this rule may apply to particular instances of meta-evaluation, there are likely an equal number of contexts or meta-evaluation models for which this guideline is inappropriate. The Standards should not engage in this level of micromanagement. (F) presumes a summative meta-evaluation, conducted in an authoritarian model, only one of many possible approaches. (See A11 Impartial Reporting for discussions of authority and reporting.) While (G) closely parallels the treatment of Formal Agreements (P2) in evaluation, a noteworthy distinction is that the client of the meta-evaluation has not been identified or discussed, making the issue of *who* determines the reporting parameters a bit more ambiguous. (H) addresses the focus of meta-evaluation, rather than procedural steps. I find this a much more fruitful avenue of argument to pursue, and I would expand (H) into several distinct guidelines so that the intersections with cultural concerns can be made visible. For example, evaluating the initial conceptualization of the evaluation design for congruence with cultural context; evaluating culturally appropriate instrumentation, including translation procedures; examining the power relationships implicit in data collection, assuring that respondents were treated with respect; insuring that all promised protections of human subjects were duly maintained in data handling; scrutinizing the culture-bound assumptions implicit in data analysis and the value frames used in data interpretation. (I) is a reasonable guideline, although it appears to be written more for formal than informal meta-evaluation.

Common Errors. (A) I agree that meta-evaluation should be incorporated in early thinking about evaluation. Ironically, the tendency of this standard to think of meta-evaluation as external and summative contributes to this error. (B) This is essentially an error of Information Scope and Selection (U3) at the meta-evaluation level. (C) As worded, I agree that this would be an error; however, I disagree that external meta-evaluation should be privileged over internal in all instances. Evaluation may be powerfully advanced and improved by thoughtful internal meta-evaluation. I don't this standard accurately reflects that potential. (D) Agree, but then it seems appropriate to take up the issue of *in*formal meta-evaluation. Should evaluators every step away from a critically reflective meta-evaluative mindset toward their work? I would say no. Reflections on cultural competence, for example, demand a meta-evaluative stance. (E) Agree. This would be a misuse of meta-evaluation. (F) This returns to the issue of Evaluator Credibility (U2) and power relationships in evaluation. Whether or not this is an error seems very much dependent upon the role, context and the purpose of the meta-evaluation.

Illustrative Case—Description. The evaluand is follow-up evaluation of corporate training courses in the telecommunication industry. The meta-evaluation is sponsored by the training Advisory Board (TAB), a committee within a 20-company consortium. This committee developed standards for course development and assigned course designers to projects. One of the standards was that follow-up evaluation be conducted to determine the post-training job performance of course graduates, but reportedly this standard was not enforced. Under criticism for the quality of its products, TAB convened a task force consisting of two evaluation specialists, one instructional technologist, and two experienced training managers to conduct a meta-evaluation. No other information is given on the characteristics or credentials of the meta-evaluators, nor is their internal/external relationship to the evaluand precisely spelled out. The charge was to document the manner in which courses were being evaluated and to recommend ways of improving the process. Specific questions addressed the extent of evaluation, compliance with quality standards, factors influencing deficiencies, and recommendations for remedial action. Data collection involved document review of all evaluation reports and interviews with project managers. The meta-evaluators identified deficiencies in the number of completed evaluations of post-training job performance and in the technical quality of the evaluations themselves. Based on the interview data suggesting causal factors behind these deficiencies, the meta-evaluators recommend seven action steps, all but one of which was reportedly implemented. Interestingly, the step that was not implemented was potentially the most impactful—hiring a full-time evaluation specialist to train, monitor, coach, and metaevaluate the work of the course developers. Allegedly, the number and quality of the evaluations improved over the next three years, though it is not entirely clear who was monitoring that improvement, the TAB or the task force. To me, it reads as if the task force competed its work with the action recommendations, leaving the TAB to monitor the implementation and impact of those recommendations. This does not inspire confidence, since that body had not previously demonstrated competence in this matter. Also, given the finding that the internal/external role of the evaluators was a significant factor in the original evaluations—evaluators who had not developed the courses they were evaluating were more likely to recommend changes than designers evaluating their own courses—the relationship of the meta-evaluators to their evaluand merits comparable attention.

No cultural context information is given in this case, though the meta-evaluators spoke of "competition among projects for limited personnel resources, the sometimes punishing consequences of evaluation, and the volatility of the course content" as factors contributing to the dearth of follow-up evaluation. These dynamics of organizational culture are not explored nor are they reflected in the action steps subsequently implemented. The one recommended action that involved personnel resources was rejected.

Illustrative Case—Analysis. The analysis lauds the case as illustrating the potential of meta-evaluation to produce change. The analyst asserts that the orientation of the meta-evaluation toward future improvement of the evaluation process mitigated defensiveness. Analysis focuses on the action steps that were implemented, but it does not question why the recommendation for additional personnel resources was not. This is especially puzzling since the TAB is praised for its strong commitment to act upon the findings, and we are told that they had the power to implement the recommendations without further approval.

The analysis adds new information about the meta-evaluators, who were chosen for "their competence and credibility." We are told that they represented different companies in the consortium, had different disciplinary backgrounds, and were not associated with any of the projects under investigation. These characteristics appear to support the credibility of the meta-evaluation, though the politics and demographics of the consortium are still unknown.

The last part of the analysis touches upon potentially relevant dimensions of organizational culture historically surrounding evaluation. The analyst correctly points out that the political pressures on the TAB were never identified and could shed important light on evaluation in the consortium. The analyst also picks up on the "signals" that management was giving about the evaluation function, and points out that the meta-evaluators did not explore the politics surrounding course development. These are good questions to raise, and they could potentially lead to clearer understanding of important dimensions of organizational culture. I think it is important to our work that culture be understood as encompassing characteristics, values and concerns of organizations and communities as well as of individuals.

◙ Add an Illustrative Case that highlights the influence of contextual factors. For example, a community that resisted the evaluation because the evaluators were perceived as not understanding or respecting the community, may reject conclusions inspire of justification offered.

Supporting Documentation.

◙ The source documentation seems especially outdated for this entry, which is puzzling since it is the newest standard. As you can see below, I have no problem with classic work being cited, but current thinking should also be reflected.

Cooksy, L. (1999). The meta-evaluand: The evaluation of project TEAMS. *American Journal of Evaluation, 20*, 123-136.

Schön, D. A. (1983). *The reflective practitioner*. New York: Basic Books.

Stevens, C., & Dial, M. (1994). What constitutes misuse? *New Directions in Program Evaluation*, No. 64, (pp. 3-13). San Francisco: Jossey-Bass.

Stufflebeam, D. (1994). Empowerment evaluation, objectivist evaluation, and evaluation standards: Where the future of evaluation should not go and where it needs to go. *Evaluation Practice, 15*, 321-338.

Stufflebeam, D. (2001). *Evaluation models, New Directions for Evaluation, No. 89*. San Francisco: Jossey-Bass.

Return to Table of Contents