Thursday, November 2, 3:20 P.M. to 4:50 P.M.

SESSION 273: Presidential Strand Panel **Room:** Lanai

Sponsored by the Topical Interest Group on State and Local Government

The Use of Evaluation in the GPRA Reporting Process - Analysis of the First Year Effort

Chair: Elmima C Johnson, National Science Foundation

Federal agencies have submitted the first reports required by the Government Performance and Results Act (GPRA), which mandates measurement of federal agency progress toward self defined performance goals. This increased attention to program accountability has led to expanded use of program evaluation methodology for the assessment of results. The performance reports offer a variety of information on program outcomes. However questions have been raised regarding the effectiveness of program evaluation in regard to data collection strategies and implied attribution of program outcomes to the investment of government resources. This panel will (1) provide an overview and analysis of the first year effort from the perspective of the General Accounting Office (GAO) and the education programs of two federal agencies (NSF and ED) (2) assess the effectiveness of the evaluation strategies and mechanisms used; (3) discuss agency expectations regarding report utilization, and (5) identify steps to strengthen the evaluation process and program outcomes.

Panelists: *Use of Program Evaluation in Federal Agencies' GPRA Performance Reports*, Elaine L Vaurio & Stephanie Shipman US General Accounting Office

In spring 2000, GAO will assess how well federal agencies' first performance reports under GPRA provide objective information on agencies' achievement of statutory objectives and the effectiveness and efficiency of federal programs and spending. Key concerns include the completeness, validity and credibility of the data presented to portray program performance. In addition, we will identify whether and how agencies used program evaluation results to address previously identified analytic challenges, such as measuring results under limited federal control, explaining why agencies did not achieve their goals, and ensuring that their performance data is credible. In this paper, we will discuss the key findings of GAO's review of performance reports from across the federal government, as well as highlight best examples of integrating program evaluation results into a balanced, objective picture of program performance.

The Use of Evaluation in GPRA: A Case Study - The Department of Education, Alan Ginsburg, US Department of Education

In this presentation, the Department of Education will describe key features of the 1999 Performance Report and the 2000 Annual Performance Plan, which they submitted to Congress under the Government Performance and Results Act. The presentation will focus on approaches to developing accurate indicators and high quality data for those indicators; use of evaluation data to provide needed information on program effectiveness, and ways in which the Department is working to improve its utilization in the agency.

GPRA at the National Science Foundation: The First Year Experience, Elmima C Johnson, National Science Foundation

Implementing GPRA has been a challenge for NSF and other agencies whose missions involve research activities. Both the substance and timing of outcomes from research and education activities are unpredictable ...long term outcomes do not lend themselves to quantitative reporting. Therefore NSF developed ....the "alternative format", which is a qualitative scale for the assessment of outcomes." (NSF GPRA Performance Report FY 1999, Executive Summary) This presentation will describe the implementation and results of the reporting process agency-wide, with a specific focus on the assessment of outcomes of science and mathematics education programs. The discussion will include a review of the management structure created, and the role of various internal and external groups in report preparation. The various reports prepared and their proposed use will be highlighted, along with projected changes in the process for FY 2000.

Mutuality of Impacts of Evaluation and GPRA at NSF, Conrad G Katzenmeyer, National Science Foundation

GPRA and evaluation have had mutual effects at NSF. GPRA has provided the rationale and motivation for encouraging new data collections and studies, while evaluations have helped to define the direction and limitations under which GPRA can operate. This presentation will discuss how GPRA has expanded evaluations across organizational units as well as within programs. In addition, the GPRA requirements have led to the initiation of evaluation activities prior to programs being implemented in several instances. On the other hand, evaluations have shaped how GPRA and program monitoring can be combined with other data to create a comprehensive view of programs. But the major issue remains the extent to which GPRA and evaluation results are used. The impact on use will be the central focus of this discussion.

SESSION 274: Alternative Format **Room:** Akaka

Sponsored by the Topical Interest Group on Business and Industry

The Role of Measurement and Evaluation in Achieving Business Results

Chair: Trude J Fawson, The Training Impact Group

Presenters: Arlene Brownell, The Training Impact Group
Trude J Fawson, The Training Impact Group
June P Maul, US West
Jennifer Unger, University of Southern California

A formal, 90 minute, alternative format presentation by four speakers who worked in unique partnership to conduct a year-long measurement and evaluation project in a business setting. The four presenters -- a business manager (the client), two external consultants, and a university-based statistician -- will describe the needs and challenges of a measurement and evaluation intervention in a business environment, the variety of methodologies used in the evaluation, the statistical analyses designed to test relationships among a variety of data sets, the research and intervention results, a cost benefit analysis of the evaluation, and the current status. Included in the presentation, will be an overview of a process flow chart showing how people in a work environment contribute to a company's long term profitability. The flow chart pinpoints where to target measurement and evaluation of performance, and can be generalized to multiple companies.

SESSION 275: Business Meeting and Presentation **Room:** Koko

Sponsored by the Topical Interest Group on Environmental Program Evaluation

TIG Chair: Emmalou Norland, The Ohio State University

Program Chair: Elizabeth J Hall, The Ohio State University

SESSION 276: Panel **Room:** Waimea

Sponsored by the Topical Interest Group on Quantitative Methods: Theory and Design

Modern Measurement Theory: Applications in Evaluation Research

Chair: Patrick E McKnight, University of Washington,

Sound measurement is important for all phases of evaluation, however, most evaluators rely on outdated and poor instruments that often fail to assess the constructs that they are most interested in measuring. Until recently, most measurement assessment was highly technical and beyond the reach of most applied scientists. Computers and inexpensive software have changed this situation. Modern measurement development and assessment procedures are becoming widespread and show great potential in helping evaluators generate or rescale measures to suit their purposes. This panel will cover several direct applications of modern measurement procedures including confirmatory factor analysis (CFA), Item Response Theory (IRT) and Rasch measurement analyses.

Panelists: *Assessing Disability Claims: An IRT Rescaling of Traditional Measures*, Patrick E McKnight, University of Washington and Miles McFall, VA Puget Sound Health Care System

Various methods and means for evaluating Veteran's disability claims have been used and discarded for untold

reasons. One thing that remains is the fact that the current methods simply do not help much in the adjudication of disability claims. Information is unsystematically weighed and included based upon the government adjudicator's preferences and claims are awarded based upon this process. The present study will illustrate the use of IRT and cablibration procedures to help determine the level of disability of veterans with post-traumatic stress disorder (PTSD). Results from this work may be used in future adjudication claims.

Rasch Analysis of Two Measures of Post-traumatic Stress Disorder, Kendon J Conrad, University of Illinois at Chicago

Rasch analysis has become widely used in education and psychology, but has rarely been applied in program evaluation. Rasch converts ordinal measures into linear, interval measures. The information to be yielded from Rasch analysis includes item and person diagnostics that reflect upon the quality of the test as well as the nature of the participants and their responses to the test. This presentation will demonstrate the application of Rasch in an analysis of two established measures of post-traumatic stress disorder, the Mississippi Scale and the Clinician-Administered PTSD Scale in a sample of 861 veterans. Implications of better measurement for the improvement of program evaluation methods and for program evaluation as a field will be discussed.

Rasch Analysis of the Worksite Health Climate Scale, Karen Conrad, Craig Velozo, Paul Reichelt & Karen Chang, University of Illinois at Chicago

The Worksite Health Climate Scale (WHCS) (Ribisl & Reischl, 1993) was developed to measure workers' perceptions of the health climate at work. Developed using classical test theory, the WHCS measures the three dimensions of organizational support, interpersonal support, and health norms. The purpose of this presentation is to demonstrate the use of Rasch analysis to improve the construct validity of the WHCS. Rasch will be used to develop an equiprecise linear measure that places persons and items on the same scale. We will demonstrate how to use Rasch to test and refine the item hierarchy, identify the most appropriate number of item response categories; and test person and item fit. The implications for program evaluation will be discussed.

Exploring Measurement Invariance with IRT and CFA, Florian Schmiedek, University of Mannheim

Measurement invariance of psychological tests is an important precondition for comparisons of scores across samples from different populations. Confirmatory factor analysis (CFA) with multiple groups allows to assess the invariance of factor structures. Item response theory (IRT) provides methods to assess the invariance of Item characteristic curves. While with CFA, multidimensional factor structures with factor intercorrelations can be compared, IRT has the advantage of allowing more detailed analyses of multicategorical items. The theoretical and practical differences of both methods are discussed and an application to data of the Test of Typical Intellectual Engagement (Goff & Ackerman) from American and German student samples is presented.

SESSION 277: Panel **Room:** Niihau

Sponsored by the Topical Interest Groups on Evaluation Use & Theories of Evaluation

Making Sense of Critical Multiplism: Pragmatic Frameworks for Designing Useful Evaluations

Chair: George Julnes, Utah State University

Introduced to the field of evaluation in 1985, Cook's critical multiplism uses the logic of triangulation to enhance inquiry. Triangulation from multiple perspectives-including multiple methods, multiple stakeholder perspectives, and multiple syntheses of impact and valuation-is offered as a way to strengthen evaluation by deepening our understanding of programs, policies, and their contexts. The promise of a commitment to multiplism, however, has remained something of a promissary note. Practicing evaluators wanting to embrace this approach have found that the encouragement to incorporate multiple perspectives has not been accompanied by useful guidance that sorts through which of the almost infinite possible multiplist alternatives might be most appropriate for particular evaluation contexts. In support of critical multiplism, this panel presents and develops a pragmatic framework that supports evaluators who need to select among multiple methods and perspectives in order to provide relevant information that warrants use by program and policy decision-makers.

Panelists: *Toward Usable Guidelines for Mixing and Making Methods Multiple*, Melvin M Mark, Pennsylvania State University

Admonitions to mix methods and practice multiplism come from numerous sources. In general, the soundness of such advice is not equaled by the explicitness of the recommendations about which aspects of methods to mix and which aspects of an inquiry to make multiple. This paper presents and discusses three levels of guidelines that evaluators and others might find useful in trying to practice mixed methods and multiplism. The first level of guidelines is quite generic and acontextual (e.g., Guideline A.1: If some aspect of inquiry can be made multiple with no practical addition of cost, do it; Guideline A.2: Mixing/multiplism is more important to the extent that a single method may lead to the wrong answer, and the wrong answer would be costly). A second set of guidelines is somewhat less generic, and more tied to the specifics of evaluation and social inquiry, but still acontextual (e.g., Guideline B.1: In general, addition of a measure is relatively low in cost, so mixed/multiple measurement of key variables should be relatively common). A third set of guidelines is more contextual (e.g., Guideline C.6: If a decision is to be made about program scale, then decisions about average effects, the range of effects, etc., are likely to be critically important, and mixing/multiplism should be mostly in service of these questions, presumably with quantitative methods in the lead). Though more contextual guidelines seem preferable, any set of contextual guidelines of limited length would probably not match the complexity and nuances of the settings in which evaluation takes place, so the more generic guidelines are important in helping guide judgment making.

Mixed Methods in Support of Program Decisions: Combining Survey Techniques To Reveal Contradictions and Commonalities, Lani Van Dusen, Utah State University

The importance of multiplism in building evaluation capacity can be seen in research that uses two very different survey techniques to evaluate the same underlying attitudinal constructs. Using these different approaches with the same sample led to dramatically different results. Namely, using the first technique, there was no demonstrated change in attitudes as a result of a 15-week intervention, while the results from the second technique showed a marked impact on attitudes of the program. This apparent contradiction in results is not surprising to practicing evaluators, but it points to the need for deliberate strategies for mixing methods to provide meaningful support for program decisions. Based on suggested resolutions in this particular study, a framework is offered to guide evaluators in using multiple methods appropriate for particular evaluation settings.

Evaluation and Audit Cross-discipline Models: A Multiplist Approach to Strengthening Information Quality, Valerie J Caracelli & Judith Droitcour, US General Accounting Office

This paper addresses one aspect of the critical multiplist framework by examining how audit and evaluation approaches as cross-discipline models provide useful information for programs and policy. GAO, and other professional services organizations focused on accountability, perform a variety of studies. Typically, these studies address questions in one or both of the following categories: (1) Audit-emphasis questions involving financial auditing, compliance, management controls, and agencies' attainment of pre-set goals and (2) evaluation-emphasis questions involving the development of new criteria, assessment of impact or net effect and other outcomes (including unintended outcomes), examination of technical soundness of government information (beyond completeness, accuracy, and consistency), and prospective analyses. Examples of cross-discipline models will be discussed and contributions to study quality emphasized. Audit and evaluation studies can bring different facets of information together, furnishing more comprehensive information on performance. To the extent possible, the paper will discuss how appropriate cross-discipline models might be selected.

Multiplism in Service of Assisted Sensemaking: Crafting Multiple Perspectives of Welfare Reform, George Julnes, Utah State University

Critical multiplism emerged as a post-modern response to the extremes of logical empiricism and radical relativism. Though the idea of strengthening evaluations through the 'triangulation' of multiple perspectives is an elegant position, critical multiplism has not been developed in ways that offer concrete guidance on which perspectives are most useful in specific evaluation contexts. Viewing evaluation as assisted sensemaking helps focus critical multiplism by (1)

emphasizing the perspectives associated with our natural capacities for making sense of our world, (2) distinguishing the different linkages in evaluative reasoning and identifying the perspectives appropriate to those linkages, and (3) linking the analytic answers from evaluation to the practical decisions of administrators and policymakers. The strengths of this view of critical multiplism are illustrated with a statewide study of welfare reform, one that uses multiple perspectives to identify mechanisms responsible for various post-welfare outcomes and to support value judgments about those outcomes.

SESSION 278: Panel **Room:** Hilo

Sponsored by the Topical Interest Group on Non-profit and Foundations Evaluation

Evaluating Prevention Programs in Their Developmental Context

Chair: Sharon G Portwood, University of Missouri at Kansas City

Given the increasing emphasis on program accountability and outcome evaluation, a growing number of non-profit agencies are seeking evaluation services. However, many of these agencies and their programs have not yet reached a point at which traditional outcome evaluation is truly useful, or even possible. This presentation will offer a framework for designing an evaluation that is sensitive to the developmental stage of the program itself (i.e., where the program is in its own "lifespan") and illustrate the application of this framework to the evaluation of three prevention programs involving youth. Each program represents a different developmental point (i.e., program "infancy," "adolescence," and "maturity"). Following the overview, individual presentations will focus on a specific evaluation to exemplify differing evaluation needs and how particular designs and methods have been employed to best respond to those needs. The session will conclude with the executive director of one program offering comments from the perspective of the evaluation "consumer."

Panelists: *A Developmental Approach to Program Evaluation*, Sharon G Portwood & Penny Ayers, University of Missouri at Kansas City

Although increasing numbers of agencies have become aware of the need for "outcome evaluation," many are not familiar with the basic requisites of such an endeavor (e.g., measurable goals, systematic data collection, valid measurement) and do not have these in place before securing an evaluator. In fact, many programs simply have not reached a stage at which outcome evaluation is informative, or even feasible. For example, if a school-based program does not know how, or even if, its program is being implemented within its "participating" schools, any measurement of "outcomes" is meaningless. In response to this dilemma, a framework for evaluation planning will be presented in which different designs are used at various phases of the program's "lifespan." Specific uses for planning, implementation, content, process, and outcome evaluation will be discussed and integrated into a cohesive framework. This framework will be offered as a promising strategy for designing a formative evaluation that not only appeals to agency personnel, but also best serves their long-term needs.

Evaluation in a Program's Infancy: The Children's Peace Pavilion Pledge School Program, Julia J Finkel, Kelly E Kinnison & Sharon G Portwood, University of Missouri at Kansas City

The Children's Peace Pavilion Pledges School Program is a social justice and peacemaking program designed for integration into elementary school curricula. In its first year of operation, the program has worked with nine diverse elementary schools to implement the program. A tenth school, a high school, is also participating by way of providing student volunteers to support the program. Evaluation activities began with the delivery of program materials to the schools and have subsequently focused on implementation, as well as the evaluation of program content. Data from this formative evaluation will be presented, along with a discussion of the implications of the data for program modification and the development of a long-term plan for program evaluation.

Evaluation in a Program's Adolescence: The STOP Violence Coalition, Robert G Waris, Sharon G Portwood, Gregor V Sarkisian, Darren McCormick & Andrew H Ward, University of Missouri at Kansas City

The STOP Violence Coalition seeks to address school violence through its Kindness Education and Bullying Prevention programs. While these programs were initiated over five years ago, the fact that there has been no systematic data collection (or contact with schools) places the program in a period of "adolescence." Despite the fact

that the program is feeling pressure to measure "outcomes," the fact that materials have been distributed to more than 400 schools (many of which have either not implemented the program or implemented it in various ways and/or to various degrees), has made initial attempts to obtain meaningful evaluation data extremely difficult. The results of a year-long effort focusing on implementation, content, and process evaluation will be presented. The many challenges faced by the evaluators and the responses developed will also be discussed, with particular attention to confounds presented by the use of multiple violence prevention programs within many schools.

Evaluation of the "Mature" Program: YouthFriends, Kelly E Kinnison, Sharon G Portwood, Robert G Waris & Andrew H Ward, University of Missouri at Kansas City

Formal mentoring programs are gaining in popularity; however, little evaluation research exists to support their effectiveness. This presentation will focus on the evaluation of YouthFriends, a school-based mentoring program. Arguably, this program is unique in terms of the large number (approximately 6,000) and variety of children it serves, the large number of schools (245) and districts (24) through which the program is operated, and the level of support it has received from the community. Nonetheless, its fundamental goal of connecting young people with caring adults is universal to all youth mentoring programs. Accordingly, data that addresses the extent to which YouthFriends is making a difference in the lives of children has value to a wide range of programs and policies. In the context of discussing roles for outcome evaluation in "mature" programs, data from youth, parents, volunteers, and school staff will be presented, and the general evaluation design discussed.

Discussant: Lisa Ashner Adkins, YouthFriends

SESSION 279: MultiPaper **Room:** Puna

Sponsored by the Topical Interest Group on Program Theory and Theory-driven Evaluation

Using Program Theory To Increase Evaluation Capacity: Lessons Learned from Two Work and Health Initiative Demonstration Programs

Chair: Stewart I Donaldson, Claremont Graduate University

The purpose of this panel is to discuss how program theory can be used to build evaluation capacity in community-based organizations and human service organizations operating in underserved communities. Panel sessions will draw upon the science and practice of conducting theory-based evaluations and lessons learned from evaluating two demonstration programs in 17 sites throughout California that are designed to assist both youth and adults in building job skills and finding employment. Each session will discuss different ways that theory has been used to enhance community efforts to evaluate their programs. Examples include using theory to clarify program goals, identify realistic and measurable program outputs and outcomes, and develop program success criteria so that evaluation efforts are tailored to specific efforts at each site. Together, panel sessions will demonstrate how this approach to increasing evaluation capacity seems most appropriate for maximizing the success of community programs.

Presenters: *Using Program Theory To Increase Evaluation Capacity*, Stewart I Donaldson & Laura E Gooler, Claremont Graduate University

This session will show how program theory can be used to aid in the development of evaluation capacity within organizations offering social service programs. A basic premise of this is that evaluations of social interventions are much more likely to be successful and more likely to be understood and used if they are developed using a theory-driven approach. The use of program theory can be used to help program stakeholders clarify their program goals, identify realistic outcomes and articulate success criteria. Program theory can be used to clarify linkages between program activities and program goals, and to identify evidence needed to document program success. In addition, theory can help stakeholders differentiate success or failure due to program implementation from the validity of program theory. Collaborative theory-building with program stakeholders further increases their understanding of and commitment to sound evaluation practice. Furthermore, it also helps increase the validity of program evaluation in a specific context and contributes to both improving programs and generating knowledge about how programs work and when they work.

Using Program Theory To Increase Evaluation Capacity in Community Technology Centers, Anna M Malsch, Cindy Gilbert, Laura E Gooler & Stewart I Donaldson, Claremont Graduate University

This session will demonstrate how program theory can be used to increase evaluation capacity in community-based organizations serving low-income populations. Lessons from evaluating 14 community technology centers throughout California will highlight the role of theory in building program stakeholders' evaluation knowledge, understanding, and skills. This session will discuss how program theory building can be used to increase community stakeholders' buy-in and commitment to sustainable evaluation practice. In particular, this session will demonstrate how theory-building can be used to reduce evaluation anxiety among stakeholders by clarifying the linkages between program theory and evaluation practice. Lessons pertaining to how program theory was used to identify and prioritize evaluation data with strong utility for community stakeholders will also be shared. Special attention will be given to the role that program theory plays in helping community organizations collect evidence needed to better monitor and improve program efforts, increase program success, and strengthen their dissemination efforts.

Using Program Theory to Increase Evaluation Capacity in Job Search Training Organizations, Laura E Gooler & Stewart I Donaldson, Claremont Graduate University

This session will describe how program theory is being used to increase both program monitoring and program evaluation capacity among 3 human service organizations offering job search assistance and training to displaced workers in diverse regional and economic settings. This session will show how program theory was used to help program stakeholders identify and design program monitoring data collection activities and program tracking efforts that support their continuous program improvement goals. This session will also demonstrate how theory was used to design measures that enabled program stakeholders to examine trainer effectiveness and demonstrate whether their training program fidelity was reached and sustained over time. In addition, this session will discuss the role of program theory in testing and refining a program's conceptual theory, and in turn, improving the program's theory and design. The value of collaborative theory-building for balancing tensions between meeting program research and program service goals will also be discussed.

Discussant: Laura E Gooler, Claremont Graduate University

SESSION 280: Business Meeting and Presentation **Room:** Kohala

Sponsored by the Topical Interest Group on Collaborative, Participatory & Empowerment Evaluation

Empowerment Evaluators Web-based Survey Findings

TIG Chair: David Fetterman, Stanford University

Presenter: David Fetterman, Stanford University

SESSION 281: MultiPaper **Room:** Kona

Sponsored by the Topical Interest Group on Human Services Evaluation

Evaluation of Welfare Reform and Related Initiatives

Chair: Christopher R Larrison, University of Georgia

Presenters: *Family Archetypes: A New Perspective on Families that Receive Welfare Benefits*, Christopher R Larrison, Larry Nackerud & Edwin A Risler, University of Georgia

A review of the scholarly literature shows that most analyses of welfare are based upon the premise that the overwhelming majority of welfare recipients receive benefits because they are single minority women who are undereducated and caring for a child either born out of wedlock or abandoned by divorce/separation. In an attempt to calibrate the accuracy of this long held stereotype, the authors surveyed a stratified random sample of TANF recipients in the state of Georgia. The resulting profile led to the identification of four types of families on the welfare roles. These four archetypes, as they are referred to in the paper, show that only some families fit the traditional stereotype, others are accessing the welfare system because of health problems, child abandonment, limited retirement assets, poor education, and fluctuating labor markets. The findings indicate that the present welfare system although successful to date is poorly prepared to handle the remaining welfare recipients.

Psychosocial Characteristics of Welfare Employment Outcomes: Longitudinal Assessment of Employment and Earnings, Peter A Neenan, Research Triangle Institute and Dennis K Orthner, University of North Carolina at Chapel Hill

Approximately 1700 participants in North Carolina's former JOBS welfare-employment initiative are profiled with respect to employment outcomes (continuous, cyclical, interrupted and indeterminate) and earnings over a three-year period following termination from the program. Participant outcomes are assessed relative to their personal and social strengths (mastery, depression level, job satisfaction and motivation, social support availability), human capital acquisition, and family characteristics. Participant characteristics data were obtained by means of baseline and multiple follow-up surveys administered while during JOBS program enrollment. Matching study participants with quarterly data from North Carolina's Unemployment Insurance (UI) wage data record system assessed participant employment and earnings outcomes. This state-level study is one of the few in which participant outcomes are tracked over a multi-year period, in contrast to most analyses of welfare-employment program participant outcomes that focus narrowly on immediate post-program experiences and typically omit consideration of participant personal, familial, and psychosocial characteristics.

Community Innovations in Welfare Avoidance: Preliminary Project Findings from the State of Texas, Dennis L Poole & Miquel Ferguson, University of Texas at Austin

This paper will present preliminary findings on 26 community innovation projects in welfare avoidance funded for two years by the Texas Department of Human Services (TDHS). A major emphasis in the initiative is state contracting with local community and faith-based organizations for provision of welfare avoidance services to TANF clients and potential TANF clients. Community innovations funded by TDHS vary greatly: transportation, case management, parenting and life skills, automotive repair, pregnancy prevention, education, Dress for Success, housing support, and other local welfare avoidance approaches. The paper will report preliminary findings in the following areas: mission and goal congruence, leadership, technical capacity, funding, community involvement, and performance.

The Georgia Welfare Reform Research Project – The Remaining TANF Recipients: A Research Based Profile, Edwin A Risler, Larry Nackerud & Christopher R. Larrison, University of Georgia

This paper reports on an in-depth evaluation and analysis of the impacts of welfare reform in a southern state. Data is presented from a stratified random sample of over 200 recipients who were interviewed in-home, was used as a basis of comparison. The comprehensive evaluation examined the impacts of welfare reform on client expectations, attitudes, and behaviors, with attention to factors (barriers - actual and perceived) associated with successful movement toward independence and self-reliance. Factors presented, among others, include employment issues, health, child care, transportation, and substance abuse. The goal of the project was to develop a profile of remaining TANF recipients and assist policy makers in identifying those individuals who will be most likely to struggle with the work requirements an time limits imposed under the federal guidelines.

Current Theories of Income Inequality and Health: Relevance for the Evaluation of Welfare to Work Programs, Eunice Rodriguez, Cornell University

In this paper we examine alternative explanations linking income inequality and health, and discus their relevance in the evaluation of welfare to work programs. Adjusting for factors including life style and risky behaviors is not sufficient to account for the observed mortality and health differences among low-income groups. Four hypothesis about the causal pathways that link socio-economic status and health are at the center stage of the discussion: a) the absolute income hypothesis, b) a relative income hypothesis with effects mediated by psycho-social mechanisms (e.g. social capital explanations), c) a relative income hypothesis with effects mediated by material resources, and d) the need to understand any residual effect between income inequality and heath considering the particular social, economic, and historic context. To assess the possible impact of current welfare to work programs on health status and well-being should be an important aspect of the evaluation process. Examples will be discussed.

SESSION 282: Business Meeting and Presentation **Room:** Ewa

Sponsored by the Topical Interest Group on Health Evaluation

Using Health Care Satisfaction Surveys To Evaluate the Clients' Perspective of Care as a Way of Building Organizational Capacity

TIG Chair Thomas H Cook, Vanderbilt University
Program Chair : Molly Engle, Oregon State University
Presenter: Thomas H Cook, Vanderbilt University

SESSION 283: Panel **Room:** Honolulu

Sponsored by the Topical Interest Group on Alcohol, Drug Abuse, and Mental Health

Documenting Information Transfer in Addiction Programs: Multiple Methods of Evaluation

Chair: Dianna L Newman, State University of New York at Albany

The purpose of this panel presentation is to provide an overview of evaluation methods used and lessons learned as a result of local and national evaluations of federally funded Addiction Technology Transfer Center activities with emphasis on dissemination of science based information to the field. These centers are charged with training service providers in all substance and mental health disciplines with new scientifically based knowledge in the field of addiction. The challenges of conducting a cross-site evaluation of technology training will be discussed in terms of methodology and measurement. The panel chair will introduce the program, the need for evaluation and an overview of the national requirements. The four presenters will emphasize different aspects of the evaluation efforts undertaken in this area and will discuss successful methodologies utilized as part of local and national efforts. The discussant will summarize findings as they relate technology transfer and to broader issues in this form of evaluation.

Panelists: *Evaluating Training in Addictions Professional Development: Cross-site Documentation*, Lloyd Goldsamt, NDRI Inc

This paper will present methods, results, and issues related to cross-site evaluation of professional development in the addiction area. Issues to be discussed include designing instruments that are applicable across sites, across types of professional development, and across time; aggregation of data across sites, and interpretation of findings that enhance utilization. The strengths and weaknesses of cross-site evaluation will be discussed within the contest of the program as will as issues pertaining to long-term documentation of change.

Systems Change in Addictions Professional Development: Issues in Documenting Impact, Lisa Reboy-Woolery, Mid-Atlantic ATTC

This paper will present a model for systemic change and the methodologies associated with it that were used to assess impact over a diverse group of projects and settings. The authors will establish the theoretical background of the model, its applied use in the addiction field, meta analysis technologies used to aggregate qualitative data, and examples of practice in the field.

Evaluating Training in Addictions Professional Development: Documenting Local Needs While Satisfying National Requirements, Craig Love, Brown University

The New England ATTC has been operating a very popular distance learning program. We have conducted our first impact study using an email based survey. Initial results indicate that leaders in the substance abuse treatment community are participating in distance learning and most (87%) claim to be actually using things learned in the distance learning programs and 87% also indicate that they have shared the information and materials from the ATTC distance learning courses. Methodology in online surveys are discussed. The implications of distance learning to effective transfer of technology in substance abuse prevention and treatment are discussed.

Documenting Systemic Change: A Model of Impact, Dianna L Newman & Jennifer Smith, State University of New York at Albany

This paper will present a model for systemic change and the methodologies associated with it that were used to assess impact over a diverse group of projects and settings. The authors will establish the theoretical background of

the model, its applied use in the addiction field, meta analysis technologies used to aggregate qualitative data, and examples of practice in the field.

Discussant: Craig Love, Brown University

SESSION 284: Roundtables **Room:** Kahuku

Quantitative Methods for Evaluation

(This session includes two 45-minute rotations of roundtables. The Host will ask the tables to rotate at 4:05.)

Host: To be announced

Roundtable A (First Rotation): Sampling Methods for Hard-to-reach Populations

Presenters: *Random Samples and Convenience Samples from the Same Population: Comparisons of Adolescent Survey Responses*, Kathleen A Bolland & John M Bolland, University of Alabama

Evaluators often use survey research, generally preferring probability samples and attempting to achieve high response rates. Little is known, however, about the degree to which random samples actually differ from convenience samples. We used both random samples and convenience samples in two annual surveys of over 1000 adolescents living in public housing, achieving response rates ranging from 70% to 90%. Preliminary analyses of responses from two samples in the first year suggest that respondents in the convenience sample reported more risk behaviors than did respondents from the random sample (70% response rate). Data analysis for the second year is not yet complete, but we will report whether that difference holds in the second year, with the larger response rate. We will conclude with a discussion of the tradeoff between achieving a large sample by combining random and convenience samples and the increased error created by doing so.

A Practical Sampling Method for Evaluating Programs for the "Hard to Reach": Venue-based Sampling of Young Men Who Have Sex With Men, Robin L Miller, University of Illinois at Chicago; Lillian S Lin, Farzana Muhib, Philip Smith & Wayne Johnson, Centers for Disease Control and Prevention; Ann Stueve, Columbia University; and Wesley Ford, Los Angeles Department of Health

Social Network Analysis methodology applies to relational data. One application is for understanding the communication structures of organizations and the positions of leaders. Identifying paths on which leaders are located helps to understand the influence leaders may have and how they may obtain informal feedback. A key step for doing network analysis is choosing a measure for a particular construct, such as leadership. In this study of 41 differentially effective elementary schools, principals were theorized to be more centrally located within the network and that this location would be different for effective and ineffective schools. There were mean differences in the leadership positions of the principals, but these differences were not significant at the p.05 level on all measures of centrality: degree, betweenness, and closeness. Sociograms did illustrate differences in the locations of the principal and those to whom the principal was connected. Network definition may also influence the results. This study explored these two avenues, the differences between the measures and the impact of network definition and describes the results for the school principals with the corresponding sociograms.

Roundtable B (First Rotation): Comparing Approaches to Data Analysis

Presenters: *A Comparison of Missing Data Treatments in Producing Factor Scores*, E Lea Witta, University of Central Florida

This study investigated the effects of use of four missing data handling techniques on factor scores produced from analysis of a survey instrument. A questionnaire containing 35 five-point Likert style questions was completed by 384 respondents. Of these, 166 (43%) questionnaires contained one or more missing responses. The missing data pattern was non-ignorable. Listwise deletion, pairwise deletion, regression, and the expectation maximization algorithm were used to treat the missing data. Resulting data was then submitted to factor analysis and factor scores obtained. Factor scores for each group defined by missing data method were then contrasted by multivariate analysis of variance. Less than 1% of the variance in scores could be explained by group($F=.218$, $df=30, 3496$, $p=1.0$). What if analyses will be conducted by reducing the number of complete responses.

Alternative Data Analysis in Evaluation Research: A Comparison between Ordinal and Traditional Statistical Approaches, Andy K Rudd & Andrew McConney, Western Oregon University

Frequently, evaluators are charged with evaluating a program that claims its program participants will score higher over time on a given variable, or score higher than a similar group that does not receive the program. These types of hypotheses or questions can be characterized as having an ordinal nature as can data from a variety of commonly selected assessment tools such as standardized tests, questionnaires, surveys, and attitude scales. Despite the ordinal nature of many program hypotheses or program data, evaluators typically choose traditional statistical approaches to data analysis that could produce spurious conclusions. Our purpose is to examine potential differences in outcomes when using ordinal analysis versus traditional statistical analysis. We suggest that an ordinal analysis may more accurately answer and analyze ordinal situations that are common in evaluation research.

Roundtable C (First Rotation): Examining Threats to Reliability and Validity

Presenters: *Be Careful What You Ask For: Unexpected Outcomes Related to Hawthorne and John Henry Effects While Evaluating a Prevention Program*, Amy R Juntunen, SPEC Associates

Lula Belle Stewart Center, a community based organization located in Detroit, MI, delivers prevention programming to preteens and parents through a grant from the Department of Health and Human Services, Office of Adolescent Pregnancy Programs (OAPP). The Prevention Project aims to: support sexual abstinence; increase knowledge of human sexuality; provide information to prevent early sexual involvement, venereal disease, and HIV/AIDS; and promote communication between youth and parents. Part of OAPP's evaluation involved pre, post and follow-up telephone interviews with parents who had participated in didactic/support groups, along with a comparison group of parents. This paper will compare the tracking and interviewing during two years of evaluation where response rates often averaged 100%. The audience will see how an expansion and modification of the evaluation from one year to the next, precipitated unexpected outcomes related to Hawthorne and John Henry effects.

A Comparison of Three Self-reporting Methods of Measuring Change in Instructional Practice, Tony C M Lam & Priscilla Bengo, University of Toronto

Inherent in the pre- and post- tests experimental design are two validity threats related to self-reporting: the "pretest sensitization bias" and the "response-shift bias." To avoid these biases and to enhance data collection efficiency, various methods of measuring change that do not require the use of pre-tests have been proposed and used in the literature. Because evaluation and research results may be affected by the particular method used to measure change, research should be conducted to examine method variance by comparing results generated by different methods. In our proposed presentation, we'll report findings from comparison of three self-reporting methods of measuring third grade teachers' change in mathematics instructional practices. These methods include the post-retrospective pretest design (reporting current practices and practices at an earlier time), the post-perceived change design (reporting current practice and the amount and direction of change), and the perceived change design (reporting the amount and direction.

Roundtable D (First Rotation): Leading Edge Techniques for Measuring Community and Social Structures

Presenters: *Visualizing Communication Networks Based on Understanding Variations in Measures and Networks*, Maryann Durland, Rockman ET AL

Social Network Analysis methodology applies to relational data. One application is for understanding the communication structures of organizations and the positions of leaders. Identifying paths on which leaders are located helps to understand the influence leaders may have and how they may obtain informal feedback. A key step for doing network analysis is choosing a measure for a particular construct, such as leadership. In this study of 41 differentially effective elementary schools, principals were theorized to be more centrally located within the network and that this location would be different for effective and ineffective schools. There were mean differences in the leadership positions of the principals, but these differences were not significant at the p.05 level on all measures of centrality: degree, betweenness, and closeness. Sociograms did illustrate differences in the locations of the principal and those to whom the principal was connected. Network definition may also influence the results. This study

explored these two avenues, the differences between the measures and the impact of network definition and describes the results for the school principals with the corresponding sociograms.

New Approaches to Neighborhood Indicator Development: Using Spatial Probability Techniques and Multi-attribute Utility Analysis To Facilitate the Evaluation of Neighborhood-based Social Service Programs, Elliott T Graham, Ramsey County Human Services Department

As interest in neighborhood-based social service programs has grown, program evaluators have become increasingly aware of the conceptual and technical obstacles to developing meaningful measures of neighborhood-level change. This paper will provide an overview of these challenges in neighborhood indicator research and will present research findings on an alternative approach to neighborhood indicator development. Specifically, the paper will first demonstrate a spatial analytical technique for defining conceptually meaningful physical boundaries for an urban neighborhood. Second, the paper will describe the application of a structured decision making model known as multi-attribute utility analysis (MAU), which facilitates the development of a longitudinal index of social, economic, and demographic change within the neighborhood boundaries defined in the first phase of the research project.The most significant implications of this research lie in its potential impact on evaluation practice, specifically, on the development of better methods for measuring and understanding the impact of community-based social service programs on urban neighborhoods.

Roundtable E (First Rotation): Validating a New Measure of Workplace Safety Climate: A Preliminary Investigation

Presenters: Brett A Magill, American Red Cross
Anthony Gallagher, American Red Cross
Christine Carr, American Red Cross

Using a theory-based approach, we created and tested a new measure of workplace safety climate--shared perceptions of members of a workplace about its safety environment (Basen-Engquist, et al., 1998). This new measure was administered to 472 participants in 21 workplaces before and after First Aid/CPR education. Factor analyses on the 12-item scale resulted in a single factor solution. The scale demonstrated acceptable internal consistency at both pre-test and post-test, with alpha estimates of .93 and .96 respectively. Empirical tests of construct validity were conducted. Based on previous research and climate theory, convergent and divergent validity of the new measure was explored. Results of this preliminary investigation suggest that the new instrument makes a valuable contribution toward measuring workplace safety climate. Limitations to test validation in the present study are discussed and implications for further research are identified.

Roundtable A (Second Rotation): Examining Survey Measurement Error

Presenters: *Adolescent Responses to Multiple Questions about Frequency and Recency of Risk Behaviors: Within-wave and Cross-wave Comparisons*, Kathleen A Bolland & John M Bolland, University of Alabama

Evaluators often use survey research to determine the frequency and recency of risk behaviors. Various item formats have been suggested to increase the accuracy of responses. Concern arises especially when respondents are young and topics are sensitive. We surveyed over 1700 adolescents in low-income housing in a metropolitan area, asking questions such as "how often did you carry a gun in the last 30 days?" and "have you ever carried a gun?" Preliminary results reveal that the response inconsistency regarding similar questions is low, but that adolescents who respond inconsistently about one risk behavior tend to respond inconsistently for other risk behaviors. We will report results of both within-wave and cross-wave analyses (e.g., did adolescents who reported in 1999 that they had never carried a gun report in 1998 that they had carried one last week?). We will provide recommendations regarding appropriate wording for questions about risk behaviors.

Errors in Survey Designs: Sources and Consequences, Shahpar Modarresi, Prince George's County Public Schools

Survey designs are appropriate tools for collecting data in the practice of program evaluation. The survey designs may provide answers to the questions raised in the outcome and/or process evaluations of complex social programs. Language plays an important role in collecting data in surveys. If survey respondents misunderstand a question in

any questionnaire, then their responses will incorporate some errors in the measurement process. As a result, the survey method is weakened and this is a major concern for many evaluators. The existence of measurement errors in survey data has the potential to distort the results of univariate, bivariate, and multivariate statistics. To reduce the sources of errors in survey designs, serious attention must be given to the development of the survey instruments. The purpose of this paper is twofold. First, to identify the sources and consequences of measurement errors in survey designs. Second, to offer practical strategies to lessen those errors.

Roundtable B (Second Rotation): Applications of Spatial Analysis

Presenters: *Incorporating a Spatial Analytical Component into an Evaluation of a County Curfew Law*, Caterina Gouvis, The Urban Institute

This paper shows how the use of a geographic information system (GIS) to examine the possible effects of a youth curfew enhances a typical evaluation model. The analysis was structured to answer the following questions with regard to the Prince George's County, Maryland curfew law: (1) Was crime-clustering in high victimization areas during curfew and non-curfew hours reduced after the curfew began? (2) Did any high crime areas completely disappear after the curfew was implemented? and (3) Were new problem or high crime areas created after the curfew began? These are valid questions with implications for allocation of community resources that would not have been answered without spatial analytical techniques. Spatial analysis, when added to traditional evaluation methodologies, adds depth to understanding by the analyst while generating 'ready to view' material for policy makers.

Implementing a Spatial Framework for Process Evaluation of Virginia's Juvenile Transfer Statutes, Sanjeev Sridharan & Heidi Vaughn, Caliber Associates; and Lynette Greenfield & Baron Blakely, Virginia Department of Juvenile Justice

We implement a spatial framework to evaluate Virginia's recent juvenile transfer statutes. The process evaluation is designed to study the linkages between local community contexts, individual level factors and juvenile transfer practices. The data from the project is integrated within a Geographical Information System. Information from twelve sites are integrated with interviews with court service directors, judges, prosecutors and probation officers. We study the regional variations in both the needs and the resources available to implement such policies across Virginia. Our analysis framework examines the linkages between macro-county level factors (such as existing levels of juvenile violent crimes), individual-level factors and juvenile transfer practices. This study is integrated with Virginia's Inmate Population Consensus Forecasting process.

Roundtable C (Second Rotation): Applications of Multi-level Analysis and Multilevel Modeling

Presenters: *Using Multi-level Analysis To Maximize the Findings of a Longitudinal Study*, Stacie A Hudgens & Peggy Glider, University of Arizona

This paper describes the efforts of a two-year longitudinal study of college freshmen's alcohol and other drug use (AOD) on campus. These efforts include multilevel analysis of survey data, aimed at measuring the effects of a normative media campaign on college freshmen's perceptions and behaviors regarding their own and others AOD use. Internal evaluators were interested in uncovering program effects that would traditionally be undetected using basic level analyses. Some methods employed in the program evaluation were item response theory, inferential statistics, and structural equation modeling. This study represents the capacity of an internal evaluation team, along with program staff, to maximize the findings through multilevel analyses.

Estimating Individual Outcomes in the Presence of Intra-county Correlation: The Use of Multilevel Modeling in Evaluating Child Welfare Reforms, Judith B Wildfire & Charles L Usher, University of North Carolina at Chapel Hill

Under waivers of federal regulations and with financial support from national foundations, many states and localities have attempted to reform child welfare policy and practice. The central question for evaluators of these initiatives is whether the changes in local policies and practice improve outcomes for children served by the system. To assess improvements in outcomes within a state, the analysis must account for intra-county correlations involving: (1) characteristics of individual children; (2) characteristics of the services they receive; and (3) aggregate characteristics

of the counties in which they live. By comparing the results of multilevel and single-level models, we illustrate how the use of analytical software that accounts for correlated data can lead to different conclusions about the effectiveness of child welfare initiatives. In addition to its theoretical significance for evaluators, the paper also has practical implications in that it reveals important limitations of widely used statistical analysis packages.

Roundtable D (Second Rotation): Heeding the Call for Quantitative Data

Presenters: *Evaluating Student Learning in the Visual Arts: The Quest for Quantitative Data*, Tracie E Costantino, Art Resources in Teaching

This paper presents the efforts by a non-profit arts education organization to respond to demands by donors and the educational community for outcome data (preferably quantitative) related to student learning in artist-in-residency programs in urban public schools. The paper focuses on the use of a pretest-posttest quasi-experimental design to assess content knowledge and the development of higher order thinking skills. Detailing the evolution of the data collection methodology over three years in three different residency programs, the paper examines the relative effectiveness of each design for reducing threats to internal validity within the highly uncontrolled environment of a public school. Issues surrounding the use of constructed response assessments to evaluate arts learning in low performing urban schools, with little arts education, are also addressed. Finally, the paper considers whether the role of the program director as internal evaluator is a sustainable method for building capacity for evaluation within the organization.

Using the Angoff Standard-setting Method in Program Evaluation: A Review, With Implications for Low-stakes Studies, Paul R Brandon, University of Hawaii at Manoa

Low-stakes K-12 program evaluations that are limited to studying outcomes with posttest-only designs must have methods for deciding if the outcomes show that the programs performed meritoriously. For this purpose, evaluators can adopt test standard-setting methods. The findings of studies using these methods will not be conclusive but will be the strongest that can be produced with the designs. This paper reviews the literature on the Angoff method, which is the most widely studied and extensively used standard-setting method. The appropriateness of using the method for low-stakes K-12 program evaluations is discussed, and each of the major Angoff steps is reviewed and evaluated. Some deficiencies in the method (and in the research on the method) that have not been thoroughly addressed in previous reviews--particularly in the step in which the description of the target standard is developed and in the step in which judges' item estimates are discussed--are addressed.

SESSION 285: Business Meeting and Presentation **Room:** Oahu

Sponsored by the Topical Interest Group on Cluster, Multi-site and Multi-level Evaluation

An Exploration of the Learning Environment at the Intersection of Cluster and Multi-site Evaluation

TIG Chairs: Beverly A Parsons, InSites
J Fred Springer, Evaluation Management and Training Associates Inc
Program Chair: Cynthia C Phillips, Third Sector Strategies
Presenter: Cynthia C Phillips, Third Sector Strategies

SESSION 286: Business Meeting and Presentation **Room:** Waialua

Sponsored by the Topical Interest Group on Pre-K-12 Educational Evaluation

Idea Exchange: Fostering Capacity for Evaluation in PreK-12 Education

TIG Chair: Maria D Whitsett, Austin Independent School District
Program Chairs: Sally L Bond, The Program Evaluation Group
Jean A King, University of Minnesota
Facilitator: Jean A King, University of Minnesota

SESSION 287: MultiPaper **Room:** Waianae

Sponsored by the Topical Interest Group on Qualitative Methods

The Variety of Qualitative Experiences

Chair: Sue E Mutchler, Southwest Educational Development Laboratory

Presenters: *Using Online, Interactive Chats for Qualitative Data Collection: The Personal Interview Revisited*, Michal Galin, Independent Consultant

There is not much research describing the experience of conducting interviews synchronously with the help of a computer, an Internet connection, and chat software. The experience of collecting interview data for an evaluation involving one internationally dispersed non-for-profit organization will be described and discussed with particular attention to what was done and how it was experienced by both respondent and interviewer. The paper will begin by revisiting the rationale for conducting personal interviews for data collection and will then discuss the positive and negative experiences of online interviewing in contrast to more traditional means of interviewing. Issues such as data quality and interview administration will be explored.

The Wisdom of Delphi: Generating a Combination of Qualitative and Quantitative Data, Bettina Greimel, Vienna University of Economics and Business Administration

The principles of the Delphi technique - named after the oracle at Delphi in ancient Greek that is well known for its wise decisions and advice - apply perfectly to the basic idea of methodological triangulation. Consisting of several rounds of inquiry of the same target group, the communications process of a Delphi study combines the strengths of group discussion and of interviewing single persons.The results of each round serve as the basis of the next round: The respondents receive feedback on the groups' overall opinion and have the opportunity to reevaluate their original answers and to answer additional questions. This procedure enables the researcher to gain a combination of qualitative and quantitative data by proceeding from open-ended to closed questions. Therefore, the Delphi technique helps the researcher deepen his or her insight into the evaluand and relevant evaluation criteria as two application examples will illustrate.

Evaluators' Perceptions of Role in Corporate Settings, Sharon Marie May, Compaq Computer Corporation

The complex role of internal evaluator in major corporations offers may challenges and rewards to professionals in the field. Not able to apply full-blown, classic program evaluation methodology, but striving for more rigor than traditional smile-sheet feedback, today's evaluation professionals in high tech organizations are striving to carve out a place for themselves in the field. The paper presents preliminary results of a qualitative inquiry into how these individuals view their work, their place within their organizations, and the impact they are having on the productivity and profitability of their companies. Also explored, is the applicability of current evaluation theories and methodologies in their daily work and specific obstacles to their increased effectiveness as evaluation professionals.

Using Qualitative Research To Identify Systemic Level Effects of a Policy Change Process, Sue E Mutchler & Diane T Pan, Southwest Educational Development Laboratory

This paper disusses how qualitative methods offer a path to a deeper understanding of the potential longer-term effects of policy-relevant activity. Presenters will describe a research project that examined the impact on state legislators of participating in community-based study circles, or deliberative dialogue programs, on education. The iterative collection of data through open-ended, key informant interviews allowed us to identify a complex set of impacts of study circles on individual policymakers. Further, this approach probed the basic philosophical, strategic, and practical beliefs that drive their interpretation and use of information from the public. Evaluations such as this, that reveal the values policymakers associate with fundamental issues around their work, allow us to look beyond the immediate effects of specific policy-relevant activities. In this case, the results advance our understanding of the potential role of both policymakers and members of the public as change agents in decision making for education.

SESSION 288: Think Tank **Room:** Molokai

Sponsored by the Topical Interest Group on International and Cross-cultural Evaluation

Strengthening the International Networks of Professional Evaluation Associations: Progress Report and Future Directions

Moderator: Arnold Love, Independent Consultant

Participants: Jean-René Bibeau, Canadian Evaluation Society
Penelope J Hawkins, Australasian Evaluation Society

Alexey I Kuzmin, Evaluation Network (Russia)
David Nevo, Israeli Association for Programme Evaluation
Donna M Mertens, American Evaluation Association
James Mugaju, Reseau Ruandais de Suivi et Evaluation
Mahesh S Patel, African Evaluation Association
Craig Russon, Western Michigan University

On 18-20 February, 2000, a W. K. Kellogg Foundation-sponsored meeting of 15 evaluation associations was held in Barbados, West Indies. The Barbados meeting led to the development of a draft framework for cooperation among regional and national evaluation organizations. This session will further the discourse by presenting the essentials of the draft framework, hearing brief progress reports from the representatives of evaluation organizations, and then discussing suggestions from all session participants about practical activities for strengthening professional evaluation associations through greater international cooperation.

SESSION 289: Panel **Room:** Kauai

Sponsored by the AEA Diversity Committee

Evaluators' Work in Cultural, Ethnic, and Linguistically Diverse Communities: Are Additional Standards and Principles Needed?

Chair: Edith P Thomas, United States Agriculture Department

The evaluation community has recognized the need to provide guidance to evaluators working often in complex social settings. Besides sufficient training, it is recognized that decisions, actions, and inferences on the part of evaluators should be informed by established codes of ethical conduct and standards of professional practice. The Joint Committee on Standards for Educational Evaluation (1994) and the Guiding Principles for Evaluators developed by the American Evaluation Association (1995) are examples of documents that reflect the culmination of reflective work by professional organizations that are concerned about the manner in which evaluations are performed. Similarly, the Joint Committee on Testing Standards have recently published the new edition of its testing and assessment standards for psychologists, evaluators, educators, and other workers (Joint Committee on Testing Standards, 1999). Despite these initiatives, however, there is concern among many professionals that more is needed when professionals are working in ethnic, linguistic, and culturally diverse communities. The proposed session will examine the general question: Are additional standards or guiding principles needed to guide evaluators' work in ethnic, linguistic, and culturally diverse populations? A subset of questions would include: Are our existing standards and guiding principles sufficient? Should they be modified to accommodate issues of diversity? Are the perspectives and issues sufficiently unique to warrant the development of a separate set of guidelines to augment existing ones? The presenters include individuals who participated in the development of the existing Guiding Principles and Standards, and evaluators who have extensive experience working in ethnic, linguistic, and culturally diverse communities.

Panelists: *Ethnic, Linquistic, and Cultural Diversity: From the Perspective of the Standards for Program Evaluation*, James R Sanders, Western Michigan University

This paper will examine the issue of addressing the need for quality evaluation in ethnic, linguistic, and culturally diverse settings through the lens of the existing Standards for Evaluations of Educational Programs.

The AEA Guiding Principles and Diversity: Reflections from the Task Force Chair, William R Shadish, University of Memphis

The primary clause in the AEA Guiding Principles that addresses diversity is III.D.5, though other clauses obviously apply in varying degrees (e.g., Section II, Clause I.). The principle is phrased at a general level because the 1994 Task Force was specifically charged to develop principles rather than standards, that is, more general statements of principle rather than more specific standards for how those principles should be implemented. The presenter will reflect on strategies for addressing the issue of diversity within the parameters of the existing principles and standards

Ethnic, Linquistic, and Cultural Diversity: The Role of Guidelines from the Perspective of Community-based Program Evaluation, Anna Marie Madison, University of Massachusetts at Boston

Evaluators working in community-based programs are acutely aware of the critical need for multicultural competence to successfully conduct their work. They recognize that the programs they evaluate have cultural dimensions that cannot be ignored, and that must be fully understood. The presenter will discuss the existing guiding principles and program standards in the context of community-based program evaluations, and provide recommendations for addressing this crucial area.

Standards and Guiding Principles: From the Perspective of Multicultural Validity, Karen E Kirkhart, Syracuse University

One of the underlying assumptions for establishing principles and standards in program evaluation is that these structures increase the likelihood that appropriate inferences will be drawn from this mode of empirical inquiry. When evaluations are conducted in ethnic, linquistic, and culturally diverse communities, traditional perspectives for examining evaluation-based inferences may be insufficient. The presenter will examine the scope and relevance of the existing guiding principles and program standards from the perspective of multicultural validity.

Discussants: Charles L Thomas, George Mason University
Juan Martinez, Community Human Services Department