The LEWIN GROUP
an INGENIX company

# A Framework for Continuous Evaluation of Office of Refugee Resettlement Formula Programs Supporting Employability Services

*Prepared for:*
**U.S. Department of Health and Human Services**
**Office of Refugee Resettlement**

*Prepared by:*
Demetra Smith Nightingale
Johns Hopkins University

March 2008

This report is part of the Lewin Group's Evaluation of ORR Programs. Opinions expressed are those of the author and do not represent official positions of ORR, the Lewin Group, or Johns Hopkins University.

## Contents

## Acknowledgments

This report benefited from the contributions of several persons at the Office of Refugee Resettlement and other offices in the U.S. Department of Health and Human Services, and from the Lewin Group Evaluation team. Pamela Green-Smith and Susan Kyle of ORR provided valuable guidance, data, reports, and knowledge throughout. They reviewed several drafts of the report and their suggestions and comments are gratefully acknowledged. Martha Newton, Nguyen Van Hanh, Josh Trent, and Ken Tota provided important conceptual and policy context, and Timothy Forbes, Gayle Smith, April Young, and James Perlmutter participated in working sessions during various phases of the study, helping us understand some of the more complex aspects of the rules and regulations regarding refugees. In addition to these officials, Emily Ball and Moushumi Beltangady reviewed drafts of the report and provided their expert comments. And Richard Jakopic, who passed away last year, is fondly remembered for his advice and input on this and many other studies.

Thanks also go to other researchers on the Lewin Group's evaluation, under which this report was produced. Their contributions were essential to this study. In particular, Mary Farrell and Michael Fishman, both of the Lewin Group, provided critical input throughout, particularly on conceptual issues, and they carefully reviewed drafts and provided useful comments that improved the report. Data analysis conducted by Bret Barden and Michael Mueller of the Lewin Group contributed importantly to several sections. Burt Barnow of Johns Hopkins University reviewed and provided helpful suggestions on the evaluation design section. In addition to Mary Farrell and Mike Fishman, Nancy Pindus and Randy Capps of the Urban Institute and Sam Elkin of the Lewin Group shared important experience about program management and operations and their insights from the field form the basis for examples used in this report.

## I. INTRODUCTION AND PURPOSE

The Refugee Social Service (RSS) and Targeted Assistance Formula Grant (TAG) Programs provide services to refugees, asylees, Cuban/Haitian entrants, Amerasians, and victims of a severe form of trafficking with the objective of helping them achieve economic self-sufficiency soon after entering the United States. The Office of Refugee Resettlement (ORR) within the U.S. Department of Health and Human Services (DHHS) administers these programs and sponsored an evaluation to assess how program services are delivered and how refugees who receive these services fare over time. The Lewin Group and its partners, the Urban Institute, Johns Hopkins University, National Opinion Research Center (NORC), and Southeast Asia Resource Action Center (SEARAC) conducted the evaluation focusing on three sites: Houston, Texas; Miami, Florida; and Sacramento, California. This Continuous Evaluation report serves as an extension of the evaluation of the RSS and TAG programs, outlining ways that ORR can better plan for and institutionalize evaluation and accountability throughout the range of refugee resettlement programs. The intent of this paper is to present a range of options ORR might consider to complement existing performance management and evaluation strategies.

The federal government determines how many refugees will be admitted to the United States. From FY 2002 to FY 2007, the proposed annual ceiling was set at 70,000, and the FY 2008 ceiling will be set at 80,000.[1] In the past five years, a total of between 27,000 and 75,000 persons have been admitted annually as refugees. In addition, certain Cuban and Haitian entrants (approximately 25,000 annually) are allowed to enter the country directly (e.g. parolees, asylum seekers). Other populations eligible for ORR-funded services include asylees (approximately 24,000 a year), certain Amerasians, and victims of a severe form of trafficking (up to 1,000 a year). For ease of reference, this report generally uses the term "refugees" to refer to all such groups that qualify for ORR services, except where delineation is necessary.

To assist refugees settle in the United States, ORR funds a number of programs that provide economic support, social services, health services, and employability services designed to aid individuals and families achieve rapid economic self-sufficiency. Most of the actual service delivery occurs through a broad network of providers including state and local agencies, mutual assistance associations (MAAs), and voluntary agencies (Volags) that have established relationships with the Department of State for reception and placement services to refugees.

The largest federal refugee programs in terms of funding are the Refugee Cash Assistance (RCA) ($36.5 million allocated in FY 2006) and Refugee Medical Assistance (RMA) ($82.0 million allocated in FY 2006) programs, whereby the federal government reimburses states for the costs of cash and medical benefits. In addition, ORR allocates funds according to Congressionally-established formulas to states and localities to provide a broad range of employability services to help individuals obtain employment and achieve economic self-sufficiency and social adjustment as quickly as possible. RSS funding for FY 2006 was about $83.4 million and the TAG funding was about $43.7 million; with a small percentage of the total RSS and TAG funds made available for discretionary allocation by ORR.

[1] U.S. Department of State, U.S. Department of Homeland Security, and U.S. Department of Health and Human Services, "Proposed Refugee Admissions Report to Congress" for each fiscal year.

RSS- and TAG-funded services are specifically intended to improve economic self-sufficiency and social adjustment, primarily through employability and support services. ORR has established extensive policies to monitor the results and performance of its programs in keeping with the regulations emanating from the Government Performance and Results Act (GPRA), the President's Management Agenda, and the Program Assessment and Rating Tool (PART). However, there has been less focus on making evaluation, on a broader scale, a part of the refugee program. There have been very few internal or independent evaluations of ORR programs, and those that have been conducted have not been adequate in rigor or scope. Strong program accountability requires solid monitoring, reporting, and evaluation strategies.

Therefore, as part of ORR's ongoing efforts to improve performance management strategies, this paper provides a framework for ORR to consider for continuously evaluating RSS- and TAG-funded employability services. The discussions and options draw from information and findings from the current Lewin evaluation of ORR programs[2], review of recommendations of the Economic Self-Sufficiency Workgroup, discussions with ORR officials and staff, and review of ORR PART and GPRA materials. The Economic Self-Sufficiency Workgroup, established by ORR in 2006, consists of ORR staff, state coordinators, representatives of Wilson/Fish programs, local and national Volags, MAAs, an employment technical assistance provider, and the Department of State. The Workgroup reviewed the operating definition of self-sufficiency, and engaged in extensive discussions about current performance measures, alternative measures, reporting, timing, and other technical issues. Input from the workgroup is incorporated into ORR's recent proposed guidelines.[3] Information from the workgroup was reviewed and used in this report as well.

The intent of this paper is to present a range of options ORR might consider to complement existing performance and evaluation strategies and the proposed guidelines related to economic self-sufficiency.

[2] Mary Farrell, Bret Barden, and Mike Mueller, "The Evaluation of the Refugee Social Service (RSS) and Targeted Assistance Formula Grant (TAG) Programs: Synthesis of Findings from Three Sites," forthcoming; Randy Capps, "Houston Case Study," forthcoming; Nancy Pindus, "Miami Case Study," forthcoming; Sam Elkin "Sacramento Case Study," forthcoming.

[3] "ORR Recommendations and Proposed Reporting Requirements and Guidelines for Economic Self-Sufficiency," ORR State Letter 07-08.

## II. CONCEPTUALIZING CONTINUOUS EVALUATION

Over the past decade there has been an increasing focus in the federal government on managing for results. The Government Performance and Results Act (GPRA) requires all agencies to develop annual performance plans with clear goals, and then track progress towards goals set. Since 2001, the President's Management Agenda (PMA) further specifies that each agency focus on continuously improving five areas of management:[4] strategic management of human capital, competitive sourcing, improved financial performance, expanded electronic government, and budget and performance integration. The fifth area, budget and performance integration, expanded upon the GPRA concepts to promote improvements in defining, measuring, and monitoring performance results, to encourage continuous improvement, and to inform budget decisions with performance results. The Program Assessment and Rating Tool (PART) which supports the PMA, is intended to ultimately assess all federal programs, and is designed to focus on goals and outcomes to help evaluate programs' overall effectiveness and to improve performance over time. There are four categories of factors included in PART: program purpose and design, strategic planning, management, and results. Each year, selected programs are identified by agencies and the Office of Management and Budget (OMB) for PART assessments, and PMA Scorecards and narrative reports are prepared and made available to the public. About half of the total PART score for each program is based on the results portion of the Assessment, which uses performance and evaluation data to determine progress towards achieving annual targets and long-term efficiency objectives. PART is intended to complement GPRA by operationalizing and integrating specific planning, management, and results activities to improve performance.

Program evaluation enters into PART in several ways. First, one of the dimensions in the strategic planning section inquires whether the program has independent evaluations, and whether the program has clear performance measures, including ambitious baselines and targets. Second, the management section requires evidence of program efficiency, including having measures established and procedures in place to achieve efficiency. Third, the results section requires programs to show evidence of continuous progress towards achieving goals, efficiency, and performance.

Thus, both GPRA and PART require agencies to have systems in place that each program can use for establishing goals and targets, measuring performance and results, and continuously tracking progress over time towards improved results. Program management through program planning processes and management information reporting are critical for conducting these assessments at the aggregate level, and program evaluation activities help establish a baseline of results, determine the effectiveness and efficiency of specific program activities, and track progress over time.

ORR has established extensive policies to monitor the results and performance of its programs, and uses this information to continuously improve programs, in keeping with GPRA and PART requirements. Because ORR programs are collaborative efforts among state and local agencies, Volags, and local service providers, it has been important to institutionalize an ongoing

[4] Executive Office of the President, Office of Management and Budget, *The President's Management Agenda, Fiscal Year 2002* (no date). http://www.whitehouse.gov/omb/budintegration/pma_index.html.

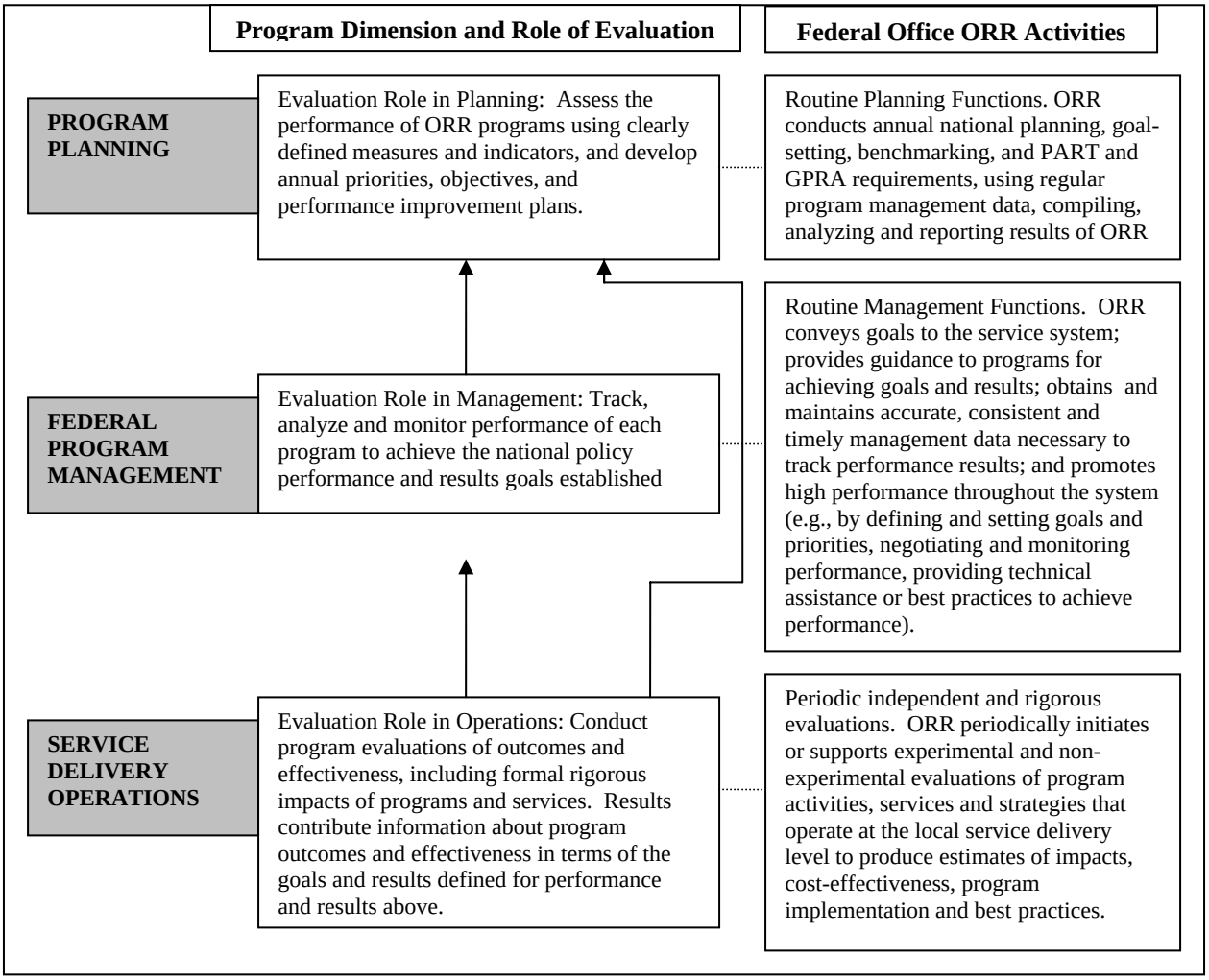
performance management and monitoring strategy into program management at the federal level and, operationally, at the state, local, and service delivery levels.

However, there has been less focus on making evaluation, on a broader scale, an integral part of the refugee program. There have been very few internal or independent evaluations of ORR programs, and those that have been conducted have not been adequate in rigor, regularity, or scope. True program accountability requires this type of regular evaluation, in addition to solid monitoring and reporting practices. This report provides a framework for more intentionally integrating the concept of evaluation into the refugee resettlement system.

In other words, the nature and types of ORR programs, along with the PART and GPRA requirements to which ORR is held, suggests a three-pronged conceptual framework for continuous evaluation, shown graphically on Exhibit 1. The general parameters of this framework are already in place, representing the confluence of ORR's programmatic management and oversight responsibilities, its broad mission of successfully resettling refugees, and its obligations to achieve results at the national level.

Two obvious assumptions underlie the framework. First, state performance is important in the context of PART; how states achieve their goals affects the achievement of ORR's PART goals. Second, evaluation is viewed as an integral part of overall program performance management, not as a separate activity, which is discussed in Section III.A. Formal evaluations of program outcomes and impacts are designed according to accepted empirical criteria and following standard methodologies, which are briefly described in Section III. B.

## Exhibit 1. Framework for Institutionalizing Continuous Program Assessment and Evaluation into ORR Programs and the Role of the Federal Office



## III. IMPROVING ORR POLICIES FOR EVALUATING RESULTS

This section presents a number of options for ORR to consider to for evaluating performance management and program results. The options discussed are summarized in the following chart.

### Summary of Continuous Evaluation Options for ORR to Consider

| OPTIONS | EVALUATION RESPONSIBILITY | RELATIVE RESOURCES REQUIRED |
|---|---|---|
| **PERFORMANCE MANAGEMENT EVALUATION** | | |
| Analysis of Current ORR Performance Measures and Relationship to Labor Market Conditions | • Expert evaluation contractor<br>• Subsequent analysis by ORR analysts | Initial investment relatively high ($300-400K), with more modest recalibration periodically (e.g., 5 years) |
| Improve and/or Supplement ORR Annual Survey | • Survey: outside contractor modifies design to improve response rate.<br>• Supplement with other surveys: ORR analysts.<br>• Establish NDNH arrangements: ORR analysts. | Relatively high |
| Establish a Regular ORR Program Characteristics Report | • ORR analysts, possibly with external evaluator assistance in initial year | Relatively low |
| **PROGRAM EVALUATIONS** | | |
| Experimental Tests Of Potentially Promising Strategies | • External expert evaluators | Range from modest cost to high cost ($200K to $10 million) per study |
| Non-experimental Evaluations Of Strategies | • External expert evaluators | Range from low to modest cost ($100K-$1 million) per study |

## A. Performance Assessment and Management

Performance assessment involves activities necessary for ORR planning and oversight of grantee program performance. There are a number of policies in place that are undergoing continuous refinement, including reviews being conducted by the Self-Sufficiency Work Group. Experience from the Lewin Evaluation provides insight into options that might further enhance the ongoing efforts at improving performance management throughout the system.

### *1. Current ORR Performance Management*

ORR currently has in place several activities related to assessing and improving performance, mainly connected to the GPRA and PART assessment, and to periodically evaluating programs. Both can be viewed as contributing to the continuous evaluation of results. Like all federal agencies, ORR is subject to the performance management parameters set out under GPRA and the President's Management Agenda and operationalized through the PART assessment. By delineating clear and concise goals, results of each federal program can be monitored and tracked over time. The goals are intended to represent realistic expectations and priorities of each program within a broader federal government policy to ensure high performance and cost-effective programming.

To meet GPRA requirements, ORR prepares an Annual Performance Plan, which presents goals and progress towards six measures of economic self-sufficiency:

- Entered employment, full time and part time
- Federal cash assistance terminations (due to earnings)
- Federal cash assistance reduction (due to earnings)
- Entered full time employment with health benefits available
- Average hourly wage of full time entered employment
- Employment Retention 90 days after entering employment

RSS and TAG are separate ORR programs, but since both sponsor extensive employment services, the two are often considered together for ORR PART performance management purposes. They are considered separately, though, in establishing state targets. Currently, each state negotiates with ORR to establish a target for each measure and states are encouraged to set or negotiate similar targets with programs within the state. Each state negotiates with ORR specific goals for measures defined by ORR, submits quarterly reports, and, by November 15, an annual report based on unduplicated participant counts for the fiscal year. ORR expects continuous improvement from one year to the next. Currently, states have six priority measures for their annual reporting to ORR that is used for GPRA purposes.

Based on grantee reporting, ORR also reports annually to Congress (as part of the President's budget request) on several performance measures for its PART reporting, including the three priority measures: (1) entered employment, (2) wage rate, and (3) 90-day job retention. ORR sets annual and long-term targets for each of these performance measures. For example, in FY 2006, the entered employment target was 56.49 percent, and for subsequent years, the target increases by one percent over the prior year. In ORR's negotiations with states, the state targets are technically adjusted to better reflect the variation in programs and the desire to encourage improvement from one year to the next. For example, for FY 2007, states with an entered

employment rate of less than 50 percent are expected to increase their performance by at least 5 percent in the subsequent year; those with an entered employment rate higher than 50 percent are expected to improve by at least 3 percentage points the next year. Each state and grantee submits Performance Reports[5] (ORR-6 Report) to ORR, and an annual report in which annual targets are compared to actual outcomes attained for each of the measures. The report may also include a narrative explanation of factors that may have affected reaching the target (e.g., difficult labor market conditions, or hard-to-employ populations). ORR is working with states to improve the quality of data in these reports and the Annual Outcome Goal Plan.

There is recognition that the accurate measurement of performance depends on the availability of management information system (MIS) data collected and reported consistently by all parts of the program, "rolling" up reports from local service providers, local agencies, and states, to obtain national results. For GPRA, PART, and program performance management, there are some data constraints, which ORR and the Self-Sufficiency Work Group are attempting to mitigate and address. For example, ORR receives quarterly reports from states, and states receive reports from local providers, each of which report on the 6 performance measures also used for GPRA at the federal level. ORR and the state agencies routinely monitor the data collection and case file information maintained locally, but the federal reports reflect aggregate data. The aggregations make it difficult to track some possibly interesting trends such as services and outcomes for particular types of refugees, since categorical details are limited.

One way ORR supplements the quarterly and annual reports is through the annual survey of refugees, designed to allow more detailed analysis of participation, services, and outcomes nationally. However, as discussed below, the results of the survey are somewhat limited because the response rate is low (38 percent for total respondents and 29 percent for new respondents in 2006).

### *2. Options for Enhancing ORR Performance Management*

In conducting the evaluation of RSS/TAG programs, the Lewin Group acquired considerable insight that could help to further improve the current ORR performance management strategies. The evaluators have also conducted in-depth surveys and developed comprehensive data bases that could be further exploited for national performance measurement purposes. Various options are presented here for ORR to consider.

An important consideration has to do with program resources. Current ORR resources are tight, somewhat limiting the options that can be adopted in the near future. For that reason, relatively lower-cost options are discussed along with options that might require relatively more investment. There is also an important caveat to the suggestions offered here. There is extensive effort underway by ORR and the Self-Sufficiency Workgroup that is already producing changes in policies and practice in this area. The options presented are not intended to replace the Workgroup's strategic planning, but rather are suggestions that could be considered consistent with the direction in which the federal office and the Workgroup are proceeding, and using the knowledge and research base developed in the Lewin Evaluation.

[5] Performance Reports are currently submitted quarterly by states, but ORR is shifting state administered and Wilson Fish programs to trimester reporting.

**OPTION: Analysis Related To Current ORR Performance Measures**

The Lewin Group's evaluation data base is very rich and could be mined more in the future to address performance measurement issues. Perhaps most useful, the Lewin outcome findings could provide guidance in establishing standards and goals in annual negotiations with states, and for routine monitoring over time. As part of the Evaluation, analysis was conducted on the three key measures used for performance management: (1) employment; (2) wages; and (3) employment retention. Because the evaluation data base links data from the survey conducted as part of the evaluation with information from the states' quarterly earnings records, special statistical analysis could be conducted about these key outcomes for participants in the sites included in the evaluation to explore factors related to program performance.

Much relevant data are included in the evaluation reports. For example, for each of the three sites, participation in the various employability services is provided, as well as estimates of employment and earnings using quarterly earnings records maintained by each state for unemployment insurance (UI) purposes. Based on experience in prior evaluations of employment outcomes, the state UI administrative data are more comprehensive than entered employment data compiled by programs and service providers, meaning that the employment rates are likely to be higher than what is obtained from ORR program reports.

Further analysis of the Lewin research data file would be useful in helping ORR better understand program performance by analyzing the variations in these three sites at the program level, and understanding how the program reported employment rates differ from rates that use the UI quarterly reports employment. By examining the range of outcomes that programs are achieving, and the range of outcomes for particular subgroups of refugees, ORR could gain more information about general benchmarks for programs and states, and reasonable improvement expected.

Since the evaluation focused on estimating individual outcomes (rather than provider or program outcomes), further analysis would be needed to more directly address performance issues at the program level. One very useful type of statistical analysis that could be conducted would examine the relationship between outcomes and labor market conditions, or how outcomes vary for certain hard-to-serve refugees. There is much research on these types of issues for mainstream workforce development and welfare programs, and it would be extremely helpful to ORR to have such analysis as well. For example, most ORR program providers cannot anticipate in advance exactly what types of individuals they will be called upon to receive (although they often specify the groups they can accommodate), and, once referred, programs cannot choose which refugees to serve (i.e., they cannot selectively "cream" and serve only the most employable individuals). Statistical analysis could be done to analyze how programs serving different types of refugees (e.g., nation of origin, education level, and work history) perform, relative to other programs. This information might allow ORR to take into account the presence of particularly difficult groups being served when negotiating goals and assessing performance.

In fact, much research suggests that programs providing employment services should consider the effect of labor market conditions when analyzing outcomes. Programs' performance on wage rates of job placements is heavily affected by the local labor market conditions, especially

prevailing wages. Statistical analysis could examine this specifically in the context of refugee programs and, again, ORR might wish to consider adjustments to some goals in negotiations with states. Regression-based statistical analysis has been used in many employment programs, either in setting goals, or as part of the negotiation procedures (providing guidelines for adjustments or explanations of variances). The U.S. Department of Labor (DOL) allows states to present the results of statistical analysis when DOL negotiates annual performance goals. Some state workforce development agencies formally adjust local program goals to account for labor market conditions, based on the results of statistical analysis. Programs in difficult labor market areas or serving difficult populations will have a harder time reaching certain levels of outcomes, but by using statistical regression analysis, adjustments to the goals can be made.

Since the Lewin Group already has developed a research file, ORR could consider using these data for further analysis. The file includes two years (eight quarters of earnings) of data, and one might consider the feasibility of continuing to track earnings over a longer period of time, to better understand long-term outcomes of ORR programs. In addition, ORR might consider requesting additional analysis to further examine the relationships between program outcomes of interest, labor market conditions, and hard-to-serve populations, with specific attention to possible adjustments to performance goals that might be appropriate.

**OPTION: Improve or Supplement Refugee Survey**

Conducting a survey of refugees on a regular basis makes good sense from a policy perspective. However, given the low response rate for the Annual Survey of Refugees, it is not clear that the current approach is the most efficient use of limited ORR resources. ORR may wish to review the annual survey and improve or supplement it. Based on NORC's experience in conducting the refugee survey for the Lewin Group's Evaluation, the low response rate is not unexpected. Refugees are a very difficult population to reach, engage, and interview. In the Evaluation, extensive effort was needed to locate the sampled individuals and conduct in-person interviews with a substantial portion of the respondents, and to conduct the interview in several languages. This suggests that achieving a response rate considered credible (e.g., over 65%) is costly and labor intensive. A number of suggestions are offered based on the evaluation's experience.

First, if ORR retains the survey, the response rate must be increased for the results to be considered credible. ORR is required to report to Congress annually about the employment, labor force, and welfare status of refugees. The annual survey is a major source of this information, and it may, therefore, be necessary to continue with it, but revisions to the process could improve the usefulness of the results. While the budget for the current survey is not known by this author, one possible explanation for the low response may be due to the limited resources devoted to these labor-intensive activities. Adequate resources must be devoted for intensive locating efforts and in-person interviewing. Telephone and mail surveys are not likely to produce adequate response rates. One alternative that could be considered is to conduct the survey every two or three years, rather than every year, which should allow resources to be used in a more concentrated manner for each survey round.

Second, ORR may wish to explore the possibility of adding supplemental questions about refugee status to some existing Census or other data collection efforts. Even if the surveys are not conducted annually, expanded information would be very useful for policy planning

purposes and for tracking change over time using different analytic approaches. For example, refugee items could perhaps be added to the Current Population Survey and the American Community Survey. These Census surveys ask respondents about place of birth, immigration or citizenship status, entry date, whether naturalized, and other information, but they do not solicit information about refugee status. Conversations with Census Bureau officials could determine whether it is possible to add items in future surveys, and the procedures and costs of sponsoring special data items in the regular surveys or one of the special periodic Supplemental Surveys. Similarly, other organizations such as the Pew Center for Research periodically conduct surveys of immigrants, but do not routinely include survey questions about refugee status or experience. It seems worth pursuing such options and determining the costs to ORR of collaborating in future survey efforts.

Third, collaborations between ORR and other federal data programs could complement the refugee survey efforts by providing routine information about employment and economic characteristics. One prospect is the National Directory of New Hires (NDNH), maintained by the Office of Child Support Enforcement, also in the Administration for Children and Families. The purpose of the NDNH is to assist in locating and enforcing the collection of child support payments from non-custodial parents. Congress also allows other agencies and researchers to access the NDNH for specific purposes, as long as the purpose is related to either Part A (TANF) or Part D (Child Support Enforcement) of the Social Security Act.[6] The case might be made that ORR is eligible to obtain data under either the provision for special research projects or for program analysis, because ORR and TANF have common populations and participants. The information is valuable enough that it is worth pursuing the feasibility of conducting a data match for either regular program purposes or for special research projects (e.g., analyzing employment and earnings of refugees, evaluating pilots or demonstrations).

The NDNH would be particularly useful to ORR because it includes quarterly earnings records on all workers in all states, meaning it captures employment for persons who move from one state to another or who work across state lines in areas where a labor market includes jurisdictions in more than one state. While there are very strict legal and regulatory provisions related to privacy that make data sharing agreements complicated, ORR may wish to continue to attempt to establish data sharing agreements that would allow regular, or at least periodic, matching of data in the ORR or State Department refugee files with NDNH files. Such matches would produce a wealth of important data that ORR could use to meet some of ORR's Congressionally-required annual reporting as well as track employment and earnings for refugees over time. One possibility would be to produce a baseline estimate of employment and earnings for a sample of refugees, and then periodically monitor trends in continued employment and earnings progressions for this sample over time, perhaps every three or five years.

**OPTION: Establish a Regular ORR Program Characteristics Report**

ORR currently collects a considerable amount of administrative and program data that could be more systematically compiled into an Annual Program Characteristics Report. Some programs

[6] http://www.acf.hhs.gov/programs/cse/newhire/library/ndnh/background_guide.htm#who presents the latest guidelines for accessing the NDNH.

(e.g., TANF) prepare a regular report that includes tables of key characteristics of states or local grantees. Information currently collected by ORR could be regularly compiled, for example, from the ORR-6 and semi-annual performance reports that include participation and outcomes data. Preparing such reports yearly or every two or three years would serve several purposes, including providing additional tables that could be included in Annual Reports to Congress and OMB, serving as an information resource tool for policy officials and the general public, and raising the awareness of all grantees about the similarities and differences between their program and others. Key characteristics could include, for example, by state: caseload, national origin, funding levels, and employability services provided, and possibly a demographic breakdown of the refugee population served, other ORR grants the state receives, local program partners, and program activity levels.

It would be very important to carefully design the Characteristics report to be descriptive and informative, not threatening and not presented in the form of a state report card. If properly designed, grantees and service providers might find the information about all programs quite useful. Such a report could also be used by all levels of the program to communicate activities and program features to the general public and public officials. Over time, more outcome or performance data could be incorporated, if grantees and providers are involved in specifying the data included in tables.

## B. Program Evaluation

A critical aspect of a program's continuous evaluation strategy should involve periodic formal evaluations of program effects on individuals served. The results of carefully designed evaluations are useful mainly for estimating the impacts and effectiveness of particular service strategies, service components, and/or projects. Evaluation results can also help to: (a) calibrate various measures used to continuously assess overall national program performance as discussed in the previous section and for reporting under PART and GPRA; and (b) provide information and guidance for improving performance because evaluations of effects on individuals can present evidence of best practices that could be replicated.

By establishing regular plans and budgets for specific evaluation projects, ORR can gradually raise the awareness of the value of evaluation in the entire system, particularly for state and local administrators. Program administrators may be unclear about what evaluation projects actually entail and how the results can be used. As discussed below, as they ask to participate in demonstrations, help design tests to evaluate best practices or try new service strategies, they will become more comfortable with the idea. As in other federal program areas, such as welfare and workforce development, they will also realize that having strong evaluation findings from a rigorous evaluation can actually help communicate their accomplishments to funders, community leaders, and government officials, and also help identify ways to improve their own program outcomes and performance.

### *1. Background on Types of Evaluation*

Public agencies and programs are increasingly expected to show that their programs are working, are effective, and are achieving what they are intended to achieve. Program evaluations are not easy to conduct and not simple to explain. Before presenting options for ORR to consider, this section provides a brief, non-academic overview of evaluation designs.

In its responsibility for implementing PART, OMB has issued a comprehensive guide to evaluation titled *What Constitutes Strong Evidence of Program Effectiveness.* [7] In general, OMB explains that in order to obtain the strongest evidence of program effectiveness, evaluations should be independent, of "sufficient scope," high quality, unbiased, and conducted on a regular or periodic basis to answer key questions about results, performance, and outcomes. According to the guide, OMB is interested in encouraging evaluations that also contribute to the planning process and help improve program performance. OMB describes a number of types of evaluations, and explains that the best (i.e., highest quality, most unbiased, and most accurate) design for measuring program effectiveness is to use randomized controlled trials to measure impacts, but also describes acceptable non-experimental designs when random assignment is not possible.

Throughout the guide book, OMB encourages the use of *experimental evaluation designs (i.e., randomized controlled trials)* to measure *program impacts*. In its most basic form, experimental design involves the use of random assignment—that is, individuals are randomly assigned to either a treatment group where they participate (or are permitted to participate) in the particular program or receive the particular service being evaluated, or to a control group where they are not permitted to participate or receive the service. Adaptations to this basic model involve planned variation tests of services or interventions, which are described below, where individuals are randomly assigned to one type of treatment or another, but there may not be a no-treatment group. It is this random assignment process that is the key feature of an experimental design evaluation. The evaluation estimates *net program impacts*, meaning the outcomes for individuals compared to what would have happened without the service, treatment, or program being evaluated.

Experimental design is considered the "gold standard" of evaluation because it produces the most accurate and precise estimate of the impact of a program on individuals. In the hard sciences and medicine, this random trials model is the basis of most laboratory experiments. In the social sciences (including public policy and program evaluation) experimental design is considerably more complicated, mainly because it is more difficult than in a hard sciences study to control the environment in which the treatment is tested. There is no scientifically controlled laboratory: people continue with their lives outside the program, the economy goes through up and down cycles, staff and administrators continue operating their programs and interacting with other programs, and there may be staff turnover or other changes as time passes. The design of the experimental evaluation must take great care to specify the treatment, establish the procedures for making the random assignments, and then monitor and ensure the integrity of the random group assignments. Careful random assignment designs consider all these potential factors in developing the randomization procedures and understanding the differences between options available to the treatment and control groups.

[7] http://www.whitehouse.gov/omb/part/2004_program_eval.pdf

The most credible program evaluations of employment impacts in workforce development and welfare programs use experimental random assignment designs, particularly when evaluating demonstrations or pilot programs testing new service strategies. The impact of the treatment (program or service) is estimated by comparing the outcomes of the treatment group individuals to the outcomes of the control group individuals. The results estimate the added (i.e., marginal) contribution of the program to what would have happened to the individuals anyway.

Experimental evaluation methods are the best designs to use for measuring program impacts. There are times and situations, though, when a pure textbook experimental design is not possible. For example, if an entire program system is changing, or if a policy is affecting or saturating an entire community, it would not be possible to randomly assign someone to a non-treatment status, since the program could be affecting the entire area or population. In this case, evaluators might seek to identify comparison or control sites where the new program is instituted in some sites and not in others, to allow the effects to be compared. To the extent possible, control sites should be randomly selected, which would produce stronger, more precise estimates of effects than non-randomly selected comparison sites. A more common issue arises in programs, like ORR, where it may not be possible to deny services to any eligible individuals, thus making it impossible to assign individuals to a "no-treatment" control group. In these programs, experimental designs might still be appropriate, though, using variants on the textbook design to test some planned variations on specific services, as discussed below in the ORR context.

Skilled researchers can usually work with administrators and staff to design feasible random-assignment evaluations. But in some cases, it may not be possible to conduct experimental design evaluations because program administrators are so hostile to the notion of denying services that they would violate the integrity of the randomization design (e.g., by serving control groups anyway, or by providing everyone the same services rather than differentiated services for testing). In some instances, there might be legal restrictions against random assignment, as there were in the pre-welfare reform era in a few states, or legal requirements that services cannot be denied to anyone, as is the case with Job Service under the Wagner-Peyser Act. Similarly, there may be an interest, or mandate, to evaluate an ongoing program statewide or nationwide with no opportunity to randomly assign individuals because they have already participated.

When random assignment is not possible, *non-experimental evaluation designs* can be used (sometimes referred to as quasi-experimental). The objective in designing non-experimental evaluations is to come as close as possible to a random assignment design despite the lack of random assignment. There are many different approaches that are used, from simply tracking changes in outcomes for participants over time, to analyzing changes that occur from one point in time to another (pre-post analysis), to qualitative and ethnographic studies or case studies of small samples of individuals.

The highest quality non-experimental evaluations, though, involve the application of sophisticated statistical and econometric analysis to control for variations in the treatment group and the control group that might be related to the outcomes and impacts being measured. The objectives are to develop a comparison group that is equivalent to the treatment group or statistically control for the differences between the identified comparison group and the treatment group. Usually, the evaluators will do both; even the most carefully selected

comparison group requires subsequent statistical analysis to control for measurable differences between individuals in the treatment group and the comparison group.

Identifying the appropriate comparison group is a critical step in non-experimental designs to determine the "counterfactual," meaning the condition of the comparison group to which the treatment or program is compared. Comparison groups might come from a natural sorting of individuals who could choose to participate in a service but do not do so. An option often considered is to identify large pools of similar individuals from a large general data base such as a national panel survey file, a census, or a list of all persons who enroll in a public agency. Regardless of where the comparison comes from, the key is that evaluators must control statistically for as many factors as possible in both groups to maximize their ability to isolate the effects of the program. The closer the comparison group is to the program (treatment) group with the exception of participating in the program, the more precise the estimates of impacts will be.

When random assignment is not possible, it is critical that the evaluators be highly trained in the use of the most sophisticated statistical techniques. Using statistical techniques, analysts can create appropriate ways to compare nonprogram individuals to program participants, and to attempt to control for biases that occur because not all factors can be controlled for statistically. While econometric techniques are now very sophisticated, there is considerable evidence that non-experimental evaluation results are not as accurate as experimental results, and that non-experimental techniques are not able to exactly replicate the results of experimental evaluations. For that reason, a good strategy is to employ multiple alternative non-experimental analytic techniques; if the impact estimates are consistent, one can be more confident in their being correct.

Whether one is using experimental or non-experimental evaluation designs, there are typically three general components to the study: A typical program evaluation consists of three components: (1) impact or outcome analysis (to estimate effects on individuals); (2) process or implementation analysis (to document the treatment and services—usually involving fieldwork, interviews, observations, and other types of data collection in the field—which also helps interpret and understand the impact estimates); and (3) cost-effectiveness or cost-benefit analysis (where the net impact estimate is put in dollar terms and compared to the costs). Of course, each evaluation may have dozens of specific research questions of interest, various sampling strategies to ensure that there are enough observations (individuals) to address the questions, baseline and follow-up surveys and data collection to track long-term outcomes, and various special components that might be needed to address particular issues (e.g., services to special subgroups of the population).

Program evaluations can range in scale and cost from very inexpensive and short-term to very large and costly. Many of the large scale welfare and workforce development demonstrations today have evaluation contracts with budgets over $15 or even $20 million dollars and take several years to produce final results. The scope and cost of an evaluation depends on the number of sites involved, the number of individuals included (sample size), whether there is a survey (which is costly), how many follow-up waves of survey administration are included, and how long the research samples (treatment and control groups) are tracked (the longer the follow-up, the more precise the estimates of impacts).

In summary, evaluations are an important source of information about the effectiveness of programs, best practices in the field, and the likely effect of program changes or new strategies that can be tested. Programs that have employment, earnings, and similar economic goals now routinely sponsor formal rigorous evaluations of impacts on individuals. To the extent possible, experimental designs using random assignment methods (with either treatment/no-treatment groups or planned variation differential treatment groups) are desirable because those designs produce the most accurate and precise estimates of impacts. When random assignment is not possible, high quality non-experimental evaluations, carefully designed, can be used, incorporating multiple non-experimental methods of analyses to maximize confidence in the results obtained. Evaluations should be independently conducted, meaning evaluators should be outside the program, to maintain unbiased objectivity. In some demonstration situations, evaluators might assist in implementation of the service treatment or provide technical assistance to the staff about the random assignment or data collection procedures, but it is critical that evaluators maintain total separation from the program being evaluated to avoid any possible conflict of interest. Thus, carefully designed evaluations can produce credible evidence of a program's effectiveness.

### *2. Current ORR Evaluation Activity*

ORR analyzes data on program activities and services, summarizes results of the annual Survey of Refugees, and sponsors special studies on particular issues or program topics. Until recently, formal program evaluations were rarely conducted. The limited focus on formal evaluations of ORR programs is understandable given the mission and structure of ORR programs. First, all refugees must be served. It is therefore not possible to adopt traditional treatment/no-treatment random assignment evaluation methods where services would be withheld from some refugees assigned to a control group. Second, grantees cannot anticipate precisely the types and number of refugees that might enter their programs. Grantees often must adapt their programs and procedures to meet the needs of new refugee groups they receive. Traditional demonstrations require the strategies being piloted to be maintained throughout the test period, which may not always be possible in refugee programs. Third, grantees have considerable discretion and responsibility for operations, and local service providers (mainly non-governmental service organizations) often operate grant-funded programs along with other programs for refugees. The state and local discretion and programmatic diversity would make it challenging for ORR to establish a multi-state evaluation.

Recent PART reviews of ORR noted concerns about the lack of formal evaluations, which led to ORR sponsoring the Evaluation of RSS and TAG. The evaluation, included in ORR's PART update, is a major effort that involves multiple analytic components:

- a large survey of refugees in the three study sites;
- sophisticated statistical analysis of participation, services received, and outcomes using a large longitudinal data file on individuals in the three programs that integrates program data, survey data, and official state quarterly earnings and employment records; and
- a comprehensive process and implementation analysis that involves program observations, focus groups with participants, and structured interviews with state and local administrators and staff in the refugee programs and related programs.

The Lewin evaluation uses a non-experimental design and statistical analysis to analyze participation and outcomes. This study represents an important step towards the types of high quality evaluations OMB is encouraging.

### *3. Options for Enhancing ORR's Evaluation Activity*

The experiences of the evaluators in the Lewin study suggest a number of options that ORR might wish to consider to continue to make progress in terms of conducting rigorous program evaluations at the local level that are consistent with and useful for the overall performance management goals of the national program. Possible experimental design evaluations using random assignment are of particular interest, but it is important to also continue to conduct ongoing non-experimental evaluations.

**OPTION: Planned Variation Experimental Tests of Promising Strategies**

ORR could consider periodically sponsoring formal, rigorous evaluations of individual impacts of potentially promising service strategies in selected sites, using experimental random assignment designs. Addressing some questions of interest might first require ORR to determine whether legal and regulatory provisions would allow particular experimental designs. For example, determining whether refugees who have access to ORR programs have better economic outcomes than refugees who do not have such program opportunities might not be possible without a waiver from the law, since all refugees are legally entitled to ORR services.

In light of ORR's mission to serve all new refugees entering the U.S., the most appropriate types of experimental evaluations would be planned variation differential treatment tests, where program participants are randomly assigned to one or another group that receives a service in a different way; meaning there may not be a pure no-treatment control group. Studies of this type could be conducted within current regulations since they need not involve withholding any required service to some refugees. For example, the impacts of different approaches to providing employment services or supportive social services could be evaluated by varying some aspect of the service to test the relative effectiveness of different approaches to the service of interest, and randomly assigning individuals to receive one version of the service or another. The particular service strategies tested using such planned variations could be determined in various ways. The process study reports from the Lewin Evaluation describe several possible service variations or potentially promising strategies that might warrant formal evaluation:

- Vary time limits on RCA for refugees to analyze the effect that allowing longer welfare eligibility periods for refugees might have on their ability to improve job skills and obtain higher wage employment. A test of this type might require waiving the time limit on cash assistance to allow alternatives periods of eligibility to be tested. One test could compare the current RCA time limit policies, to a policy that allows at least 12 more months of eligibility. Refugees would be randomly assigned to one group or the other.
- Vary the length and model of ESL programs for refugees. English language instruction is a core component of refugee programs, but many staff indicate that the primary emphasis on rapid employment may come at the expense of improved English competency. One test could compare current ESL class durations (in whichever programs participate in the evaluation) to longer durations (e.g., compare 6 week programs to 16 week programs.

Alternative tests might formally compare ESL that is integrated into the job (workplace-based educational instruction) versus traditional classroom ESL; or compare sequential models (i.e., ESL first, then job placement) versus integrated models (ESL integrated into the workplace). A test of this sort might require collaboration with the TANF agency to consider whether current work requirement policies would have to be modified or waived for refugees during the demonstration period.

- Evaluate the impact of vocational training by testing various models. Again there is interest among program managers and staff to adopt strategies that lead to better employment options with long-term progression potential. One test could compare traditional classroom-based vocational training to on-the-job training or work-based apprenticeships. A three-way test could compare these options to a group that receives only job placement services and no training. The training could be occupational only or occupational plus ESL.

- Test various bonus strategies. Some programs provide bonuses to providers who achieve certain job placement or retention rates, and the Public Private Partnership (PPP) and Wilson/Fish options allow individual client bonuses. One test could vary the amount of the bonuses to determine whether the amount results in different outcomes. In employment programs generally and for TANF recipients, it is not unusual to see bonuses offered to individual participants, again mainly based on job entry or job retention. This would also be possible to test in selected refugee PPP and Wilson/Fish, where there could be a "no bonus" random control group.

- Test various employment and economic advancement strategies, such as subsidized employment (e.g., on-the-job training, work experience/community service jobs, tax credits to employers, wage supplements to individuals), career advancement (career ladders, lifelong training), or asset development (e.g., individual development accounts, business or microenterprise development). Tests could offer such options to one random group of refugees to determine interest or "take up rate" as well as estimating impact on employment and earnings. Since some discretionary grant programs such as individual development accounts or microenterprise development accounts, have waiting lists in some locations, there may be good candidates for random assignment evaluations.

- Test strategies for special populations that may be considered difficult to serve or that have unique barriers to employment (e.g., women, older adults, teenagers, survivors of torture, victims of human trafficking, and parents of young children).

- Tests of different program administrative arrangements. There is a growing debate within the refugee community regarding which approach bests serves refugees and increases refugees' employment and self-sufficiency: a Volag-administered approach or a publicly-administered approach. A demonstration could be conducted in states or communities interested in moving to a PPP or Wilson/Fish model to test the outcomes using the new procedures relative to the status quo. Alternatively, a demonstration could be conducted among Volags serving some refugees with the Matching Grant program, while referring others to the publicly-administered program. In both examples, participants would be randomly assigned to a program in which Volags provide integrated services and cash assistance to refugees or to the welfare agency that would provide these services and make appropriate referrals for employability services.

The above are just illustrative of the types of service strategies identified through the Lewin study that could potentially be tested in refugee programs using experimental planned variation evaluation designs and that would provide very useful information to program providers about best practices related to the standard types of performance measures in the program.

Other ideas about strategies worth testing could be solicited from the field. For example, many federal agencies solicit interest from the field through grant programs. An Evaluation Grant Solicitation could be issued by ORR, asking state and local programs to submit applicants for grants that could be used to test some special service. Grant awards could be modest, serving as an incentive to participate and try new strategies, or the awards could be fairly large, in effect funding major new activities or pilot projects. One criterion for award would be that the program agrees to cooperate with the national evaluator selected by ORR.

ORR could also modify current grant program announcements to include an evaluation component. For example, many federal grant programs build in some evaluation requirements in addition to regular reporting (e.g., a "mid-term" or "final" evaluation of accomplishments). Some grant programs include provisions in the Terms of Reference that require grantees to fully cooperate with a federal evaluator, or require or strongly encourage an independent external evaluator, or, at a minimum, require self-evaluations. In any case, the evaluations would be conducted at a project, or service delivery, level, and would include analysis of objectives, outcomes, best practices, lessons learned, and improvement strategies. External evaluations by independent researchers could be encouraged by allowing grantees to use funds for evaluation purposes.

**OPTION: Non-experimental Evaluations of Selected Strategies**

In addition to sponsoring particular experimental evaluations, ORR might also consider regularly supporting important non-experimental analyses of services and participation where experimental designs are not feasible. Such evaluations should employ the most sophisticated statistical analysis and should be conducted by highly experienced evaluators. Appropriate non-experimental designs, along the lines suggested by OMB in its Evaluation Guide, could include longitudinal tracking or, as recommended by OMB, more sophisticated statistical techniques to control for statistical bias that exists when random assignment is not possible. There are many techniques used to attempt to control for selection bias, including propensity scoring, which is often used in non-experimental evaluations.[8] The objective of non-experimental evaluation approaches is to produce estimates of impacts (i.e., the difference between the two groups) that are as precise as possible when random assignment is not used. The validity of the results depends on having reliable data and variables and accurately specifying the statistical models.

In some situations, like ORR programs, there may be legal constraints that would preclude random assignment that involves a "no-treatment" control group. Non-experimental designs are

[8] There are various propensity scoring approaches, but the basic method involves pooling treatment and comparison groups of individuals and using multivariate statistical analysis to estimate the probability of participating (in the condition being evaluated). This probability estimate is the propensity score. Members of the comparison group pool are selected on the basis of how closely their propensity score matches the propensity score of the treatment group member. In some studies, more than one match is selected for each treatment group member.

particularly appropriate in situations where random assignment does not apply, such as when evaluating programs or services retrospectively, or when evaluating the effect of a system-wide change. If experimental design is not possible or not appropriate, then non-experimental designs that include comparison groups and the application of statistical techniques to attempt to control for selection bias should be used. Non-experimental evaluations can be strengthened by:[9]

- Replication (including more sites or more individuals in the evaluation)
- Additional explanatory variables (to improve the matching or selection of comparison groups)
- Multi-method design (using more than one approach, such as time series plus regression continuity or propensity scoring; or developing multiple comparison groups)
- Sensitivity analysis (to examine how well the statistical results hold up under various statistical conditions or with subsamples)

Well-designed and specified non-experimental analyses could be used to address many of the same types of service delivery issues noted above under the discussion of experimental evaluations. The process case studies from the Lewin evaluation also suggest the following possible studies that could be evaluated non-experimentally:

- Assess the long-term employment and earnings trends of ORR participants using a large longitudinal data base with program data linked with administrative data on quarterly earnings and welfare receipt
- Identify strategies associated with high-performing programs by analyzing the relative outcomes of individuals in programs with high reported performance on selected measures to programs with low reported performance.
- Document the nature and outcomes of career ladder programs, by tracking participation and employment results of individuals in those programs compared to refugees not enrolled in those programs.
- Analyze employment and earnings and other outcomes of refugees in different policy environments (e.g., in high-welfare benefit states versus low-benefit states) to identify non-employment outcomes that are related to improved self-sufficiency (e.g., English competency, educational attainment, or family stability) and factors that may be particularly relevant to outcomes in various places.
- Examine and compare service delivery and outcomes in programs that have different administrative arrangements (e.g., publicly-administered versus Volag administered programs), including multiple examples of each type of structure to improve the validity and generalizability of results.

In both experimental and non-experimental evaluations, the richness of evaluation results is enhanced by multi-site studies. For example, it would be highly desirable to select a range of

[9] These suggestions for strengthening non-experimental evaluations are from Burt S. Barnow and Marvin Mandell, 'Strategies for Evaluating Education Interventions (When You Can't Use Random Assignment)," Presentation for the U.S. Department of Education, January 24, 2001, revised October 9, 2006.

sites that represent different administrative features, including some that operate under Wilson/Fish or PPP authority, as well as matching grant and formula grants and traditional arrangements with TANF, to analyze the effectiveness of integrated services and services funded with blended sources versus standard program resources and partnerships.

Evaluations should always be conducted by outside researchers who have no vested interest in or responsibility for the program. Either ORR could contract with an evaluator or grantees receiving special awards could be required to engage an independent contractor locally. Many federal agencies require both a local independent evaluator and a national contractor that synthesizes data and results across multiple programs involved in a particular demonstration or special initiative. Regardless of how potential experiments and study programs are selected, the basic standards of high quality designs discussed earlier should be followed.

## IV. CONCLUSIONS

The previous sections presented a number of options for ORR to consider that together would incorporate ongoing evaluations of program performance and program impacts into refugee programs. Some options, such as evaluating experimental random assignment demonstration projects, should be conducted by independent external skilled evaluators. Others, such as periodically conducting matches between ORR participation data and the NDNH could be carried out by either outside researchers or ORR analysts with appropriate quantitative analysis backgrounds. The combination of various evaluation and performance management strategies, as outlined above, have the potential to promote a culture of evaluation, accountability, and continuously improved service provision throughout the refugee resettlement program.

Resource constraints are a very important consideration for all federal agencies. Some of the options presented, such as an expanded survey, inevitably require a considerable investment of resources. Other options, such as producing a Refugee Program Characteristics Report, could be incorporated into regular federal management and monitoring activities with modest additional costs. Options could be designed in different ways, contingent on the resources available.

ORR can take leadership in the area of program evaluation by including regular evaluation plans into the budget planning process (such as through planning for grant awards, and making budget requests). Instituting some of the lower cost options discussed, such as an Annual Refugee Program Characteristics Report, is particularly feasible and would focus attention on both performance management and program outcomes. Continuing to pursue options such as accessing the NDNH data to better examine earnings and employment of ORR participants, or adding refugee data items to existing surveys by other agencies and organizations, such as the Bureau of the Census, also would be low-cost priorities.

Complementing these activities, which could primarily be carried out by ORR analysts, with one or more rigorous evaluations of a program, or program service, by an independent evaluator, could help institutionalize the concept of evaluation into the entire system. It will not usually be possible to have traditional treatment/no treatment random assignment evaluations, given the mission of ORR programs to serve all refugees. However, it is possible to conduct random assignment planned variation studies to test the relative effectiveness of providing different models of a particular service delivery strategy (e.g., compare traditional classroom vocational training with on-the-job training; compare workplace based ESL training with traditional classroom training). Randomly assigning appropriate participants to one service model or another (i.e., planned variations) would allow a strong experimental test. Non-experimental evaluations could also be done to examine the relative effectiveness of different services, using administrative program data on services received, and state earnings data, for participants in one or many programs, and applying appropriate statistical analysis to attempt to control for selection bias.