Contract No.: 233-02-0086
MPR Reference No.: 6117-100

MATHEMATICA
Policy Research, Inc.

Options for an Impact Study of the NIH Extramural Loan Repayment Programs

Draft Report

May 24, 2005

Stacy Dale
Tim Silva

Submitted to:

National Institutes of Health
Office of Loan Repayment and Scholarship
Rm. 2E30
2 Center Drive
Bethesda, MD 20892-0230

Project Officer:
Dr. Alfred Johnson

Submitted by:

Mathematica Policy Research, Inc.
600 Maryland Ave. S.W., Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director:
Tim Silva

CONTENTS

TABLES

FIGURES

Figure **Page**

As "the steward of medical and behavioral research for the nation," the National Institutes of Health's (NIH) mission is to support "science in pursuit of fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to extend healthy life and reduce the burdens of illness and disability." A great deal of the research supported by the NIH takes place not on its main campus or in its other facilities but rather at universities and research institutions across the country. Most of the support for what is termed extramural research takes the form of contracts, grants, and cooperative agreements for researchers. In recent years, however, another type of support has become available: loan repayment programs (LRPs) for scientists conducting extramural research. The Extramural LRPs aim to attract and retain highly qualified biomedical scientists in positions that allow them to conduct certain types of research or research deemed particularly important.

This report, which presents options for evaluating the impacts of the Extramural LRPs by using a quasi-experimental design,[1] is organized as follows. Section I provides an overview of the loan repayment programs. Section 2 discusses the need for an evaluation and a conceptual framework. Section III presents the proposed research questions. Section IV discusses our recommendation to compare funded applicants with nonfunded applicants. Section V describes the outcome variables and other variables of interest as well as a plan to collect needed data through both a survey and secondary sources. Section VI outlines our recommended analysis methodology, which is a regression discontinuity design. Section VII presents three options for a study of relatively early outcomes involving different samples of applicants and varying time frames for data collection and analysis. Section VIII briefly describes the product that will result from an evaluation product, and Section IX offers our conclusions. The report also includes two appendices. Appendix A presents technical details on our proposed statistical model, and Appendix B summarizes what we learned from interviews conducted with officials from eight institutes and centers (ICs) as background for this report.

Ultimately, the choice of the design option for the evaluation depends on the Office of Loan Repayment and Scholarhip's (OLRS) goals for the evaluation and its desire for different levels of information on program effectiveness. In particular, if OLRS wants separate impact estimates for individual LRPs and various subgroups, such as individuals with MDs versus individuals with PhDs, then we recommend a study that includes all applicants from the 2003 and 2004 cohorts. If separate estimates for subgroups and several LRPs are not a priority, then we recommend a study of either the 2003 cohort only (which would also produce results sooner) or the Clinical LRP only.

[1] Whenever possible, evaluators would prefer to use true experimental designs, such as those used in clinical trials, in which subjects are randomly assigned to a treatment or control group. Because of the way in which LRP applicants undergo review and selection for funding, we were told at the outset that such a design was not an option.

I. PROGRAMS TO BE EVALUATED

The evaluation will potentially pertain to all five Extramural LRPs, each promoting a particular type of research or targeting certain types of scientists:

- ***Clinical Research LRP.*** The Clinical Research LRP supports clinical research as broadly defined.[2] In Fiscal Year (FY) 2002, the program was open only to researchers with NIH grants, but that restriction was lifted the following year.

- ***Clinical Research LRP for Individuals from Disadvantaged Backgrounds.*** Available since FY 2001, the program also supports clinical research but is open only to individuals from a disadvantaged background, which is defined as a low-income family (not more than 200 percent of poverty levels).[3]

- ***Contraception and Infertility Research LRP.*** The contraception and infertility program supports research "whose long-range objective is to evaluate, treat or ameliorate conditions which result in the failure of couples to either conceive or bear young" and research "whose ultimate goal is to provide new or improved methods of preventing pregnancy." Priority goes to applicants "with a clear career focus in the specialized areas of contraception and/or infertility research over those engaging in general reproductive sciences research."

- ***Health Disparities Research LRP.*** Available since FY 2001, the health disparities program supports "basic research, clinical research, or behavioral research directly related to health disparity populations and the medically underserved." By statute, 50 percent of awards must be made to members of identified health disparity populations, currently defined to include blacks/African Americans, Hispanics/Latinos, Native Americans, Alaska Natives, Native Hawaiians, Pacific Islanders, Asian Americans, and the medically underserved.

[2] Clinical research is "patient-oriented clinical research conducted with human subjects, or research on the causes and consequences of disease in human populations involving material of human origin (such as tissue specimens and cognitive phenomena) for which an investigator or colleague directly interacts with human subjects in an outpatient or inpatient setting to clarify a problem in human physiology, pathophysiology or disease, or epidemiologic or behavioral studies, outcomes research or health services research, or developing new technologies, therapeutic interventions, or clinical trials."

[3] A sliding scale used for FY 2003 awards ranged from $17,720 for a family with one dependent to $60,840 for a family with eight dependents. Applicants can demonstrate their disadvantaged status by submitting proof of having previously received (1) federal disadvantaged assistance in school, (2) loans from either the Health Professions Student Loans or Loans for Disadvantaged Student Program, or (3) a Scholarship for Individuals with Exceptional Financial Need from the U.S. Department of Health and Human Services.

- ***Pediatric Research LRP.*** The pediatric research program supports research directly related to diseases, disorders, and other conditions in children. Before FY 2003, the program was restricted to researchers with NIH grants.

A. Program Goals

The Extramural LRPs are intended to increase the number of people undertaking certain types of research (clinical) or research in certain fields (pediatrics, health disparities, contraception and infertility).[4] (Henceforth we will refer generically to "research areas.") NIH program materials have described the LRPs as intended to help both recruit scientists into these research areas and retain them. From these perspectives, the programs would be considered successful if they draw more people into the above research areas than would have been the case without the programs or if participants continue to perform research in these areas longer than they would have without the programs.

B. Program Benefits and Commitment

The Extramural LRPs are designed to make entering and remaining in the designated research areas more financially appealing. The programs repay qualified education debt up to $35,000 per year, sending payments to participants' lenders. The benefit amount depends on a participant's level of qualified debt. Average awards made in FY 2003 ranged from about $24,700 per year for participants in the pediatric LRP to about $32,300 per year for participants in the disadvantaged clinical LRP. Given that the repayments to lenders represent taxable income to the participants, the NIH offsets the increased tax liability by making payments directly to participants' IRS tax accounts at the rate of 39 percent of the loan repayments.

In exchange for the financial benefits, participants agree to conduct research in the designated areas for at least 50 percent of their time (at least 20 hours per week based on a 40-hour work week) for an initial period of two years. Thus, in the initial two-year contract period, participants could see up to a $70,000 debt repayment. In addition, awardees can apply for a competitively awarded one- or two-year renewal, extending the financial benefits and their designated research commitment over a maximum of four years.

[4] In the case of the Clinical Research LRP for Individuals from Disadvantaged Backgrounds, the goal is to promote a particular type of research *among a particular group* of doctoral-level degree holders—those from low-income families.

C. Program Eligibility

To be eligible for the Extramural LRPs, individuals must:

- Hold a doctoral-level degree (PhD, MD, DO, DDS, DMD, PsyD, PharmD, DPM, DC, ND, or equivalent doctoral degree)[5] from an accredited institution
- Conduct qualifying research for at least 50 percent of total effort (not less than 20 hours per week based on a 40-hour week)
- Have that research funded by a domestic nonprofit or U.S. government (federal, state, or local) entity
- Not be involved in research for which funding is precluded by federal law, regulations, or U.S. Department of Health and Human Services/NIH policy
- Have qualifying educational debt equaling at least 20 percent of base pay from the institution supporting the research
- Hold U.S. citizenship (or U.S. national or permanent resident status)
- Not have a federal judgment lien against property arising from a federal debt
- Not owe an obligation of health professional service to a federal or state agency or other entity during the proposed two-year LRP contract period (unless deferrals are granted for the length of LRP service obligation)
- Not be a full-time federal employee[6]

D. Application and Selection Process

The application period opens in September of each year and closes in December. Scientists apply online through the NIH Loan Repayment Web site. They can apply for only one of the LRPs and must specify their choice at the beginning of the application process. OLRS examines applications to determine fulfillment of the basic eligibility requirements and then forwards complete applications to the Center for Scientific Review (CSR).

CSR considers the type of research applicants will be conducting and then forwards the applications to the relevant IC for review, scoring, and ranking. In particular, CSR forwards applications for the Health Disparities Research LRP and Clinical Research LRP for Individuals from Disadvantaged Backgrounds to the National Center for Minority Health and Health Disparities Research (NCMHD). CSR sends applications for the Contraception and Infertility

[5] A DVM qualifies for all LRPs except the two focused on clinical research.

[6] Part-time federal employees (20 hours per week or less) engaged in research supported by a nonfederal entity for at least 20 hours per week may be eligible.

Research LRP to the National Institute of Child Health and Human Development (NICHD). Other applications are sent to the IC that sponsors research on issues similar to that which applicants propose to address. For example, a Clinical Research LRP application from a scientist researching cancer would be sent to the National Cancer Institute (NCI). NCI would also receive applications under the Pediatric Research LRP if the applicant were involved in research on cancer affecting children.

Each IC establishes a panel of outside experts, such as academicians, to review the applications. The reviewers consider (a) the applicant's potential to pursue a research career (as indicated by, for example, the appropriateness of the applicant's previous training and experience, the suitability of the applicant's proposed research activities to foster a research career, and the applicant's apparent commitment to a research career) and (b) the quality of the overall work environment to prepare the applicant for a research career (as indicated by the availability of appropriate scientific colleagues and the quality and appropriateness of institutional resources and facilities). Two or three primary reviewers score each application on a scale of 1.0 to 5.0 (high to low). The primary reviewers lead a group discussion of the application, which is then scored according to a set of standardized criteria by all reviewers. Scores are averaged and converted to a scale of 100 to 500.[7]

IC officials then consider the rank-ordered list and make award recommendations. An IC may go down the list strictly in order and fund as many applicants as possible given the available funding. Alternatively, an IC may pass over certain applicants in order to offer loan repayment to other applicants lower on the list. For example, an applicant performing research in a particularly high-priority area might be selected over someone with a higher score but whose research is viewed as less critical. IC funding recommendations are sent to OLRS, which announces the funding decisions between June and September each year. Table 1 illustrates how LRP recipients were distributed across ICs in a recent year.

II. NEED FOR AN EVALUATION

Federal government programs are coming under increased scrutiny. In recent years, the Office of Management and Budget (OMB) has been using the Program Assessment Rating Tool (PART) to review the extent to which programs have a clear purpose, are well managed, and are achieving their goals. One question that agency officials must address on the PART is, "Do independent evaluations of sufficient scope and quality indicate that the program is effective and achieving results?" The question reflects an increased focus on solid, scientific-based evidence to ensure that taxpayer dollars are invested wisely.

[7] NCMHD's review process is slightly different. Applications deemed unworthy of consideration by the full panel do not receive scores on the 100-to-500 scale.

TABLE 1

NUMBER OF NEW AWARDS, 2003

	Loan Repayment Program				
Institute/Center	Clinical Research	Clinical Research for Individuals from Disadvantaged Backgrounds	Contraception and Infertility Research	Health Disparities Research	Pediatric Research
National Center for Complementary and Alternative Medicine (NCCAM)	6				
National Center on Minority Health and Health Disparities (NCMHD)		21		106	
National Cancer Institute (NCI)	108				30
National Center for Research Resources (NCRR)	35				8
National Eye Institute (NEI)	21				11
National Human Genome Research Institute (NHGRI)	3				1
National Heart, Lung, and Blood Institute (NHLBI)	83				36
National Institute on Aging (NIA)	45				1
National Institute on Alcohol Abuse and Alcoholism (NIAAA)	32				9
National Institute of Allergy and Infectious Diseases (NIAID)	65				38
National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS)	26				6
National Institute of Child Health and Human Development (NICHD)	38		10		50
National Institute on Drug Abuse (NIDA)	59				14
National Institute on Deafness and Other Communication Disorders (NIDCD)	19				9
National Institute of Dental and Craniofacial Research (NIDCR)	8				3
National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)	36				34
National Institute of Environmental Health Sciences (NIEHS)	4				9
National Institute of General Medical Sciences (NIGMS)	8				3
National Institute of Mental Health (NIMH)	97				21
National Institute of Neurological Disorders and Stroke (NINDS)	32				16
National Institute of Nursing Research (NINR)	1				1
National Library of Medicine (NLM)	3				

By commissioning the development of a research design to evaluate the effectiveness of the Extramural LRPs, NIH officials have demonstrated a clear interest in determining whether the agency's investment is yielding its intended impact. While just a few years old, the Extramural LRPs represent a substantial public investment that has increased rapidly. Contracts executed in FY 2003 totaled just over $63 million. Furthermore, because the NIH has already initiated an evaluation of its Intramural LRPs, which attempt to attract and retain scientists in research positions at the NIH, an examination of the impacts of the Extramural LRPs will provide policymakers with a more comprehensive perspective on the general strategy of using loan repayment to increase the biomedical researcher workforce.

A. Type and Purpose of Evaluation

Mathematica Policy Research, Inc. (MPR), was asked to develop one or more options for rigorously evaluating the impacts of the Extramural LRPs by using quasi-experimental methods. The data we propose to collect for the impact study will enable us (1) to describe the role that the Extramural LRPs and other factors played in applicants' decisions to take certain research jobs and (2) to document the career outcomes of recent LRP applicants (both awardees and those not funded). The impact evaluation will measure differences in career outcomes between LRP participants and a comparison group of biomedical researchers. It will tell us what participants' outcomes would have been if they had not been awarded a loan repayment contract. That is, the impact evaluation will assess the overall effectiveness of the Extramural LRPs as a strategy for increasing the number of scientists conducting specified types of research.

The objective of conducting such an evaluation is to provide the NIH with the critical information it needs to judge whether the Extramural LRPs are producing their intended effects. On the issue of potential recruitment effects, the study will be able to highlight when applicants learned of the LRPs relative to taking an eligible position and the role the programs played in applicants' career decisions. On the issue of retention effects, the evaluation will determine whether participants continue to undertake targeted research longer than similar nonparticipants. As discussed later, the study can also address other potentially important career outcomes. Data on the relationship between the review scores assigned to applicants and those individuals' outcomes will highlight the extent to which the review criteria and process accurately predict individuals' potential for pursuing a research career. More generally, the study will increase the NIH's knowledge about the potential effectiveness of loan repayment as a strategy for influencing the labor supply of scientists engaged in biomedical research.

B. Conceptual Framework

We describe below how the impact evaluation will be guided by certain perspectives on the relationships between important variables and ideas on how the programs might lead participants to achieve desired outcomes. We focus not only on recruitment and retention goals but also on research productivity, an outcome mentioned by several of the NIH officials we interviewed and that we propose to measure in a various ways described later.

The LRPs do not represent a complex intervention. Unlike many social programs, for example, loan repayment does not involve a variety of distinct services that address participants'

several perceived needs. The NIH does not provide LRP participants with, for instance, job search or job placement assistance, coaching on how to apply for grants, or tips on how to write for research journals—all of which might help them remain in research longer or function as more productive researchers. Rather, the programs seek to achieve their goals through a single service: increased income dedicated specifically to debt reduction. This simple approach implies that a study of program impacts need not be guided by a complex conceptual framework.

With respect to the programs' potential recruitment effects, the relevant conceptual question is, How might the existence of the Extramural LRPs "cause" more scientists to enter the targeted research areas than would be the case if the programs did not exist?[8] For a given LRP to affect job choice, several conditions must apply. An individual must meet the eligibility criteria pertaining to citizenship and education and must expect to meet the debt level requirement based on an expected base salary level. The individual must know about a program before deciding to pursue a qualified position. The individual must find a position that would qualify him/her for loan repayment. The individual must take the qualified position based at least in part on the possibility of receiving loan repayment, with loan repayment a motivating factor. And, ideally, the individual must have a fairly clear sense of how much debt relief he/she would receive through a program. Thus, to study potential recruitment effects, the evaluation must determine what program applicants knew and when they knew it. To shed additional light on applicant decisions to take qualified research positions, the evaluation could query applicants for their views on the pecuniary factors (e.g., how income and benefits compared with expected living expenses) and nonpecuniary factors (e.g., job responsibilities, prestige, colleagues, location, and so on) that they considered and the importance they attached to these factors.

With respect to the programs' potential retention effects—in either the target research areas or any type of research—the major conceptual question is, How might acceptance into the LRP "cause" participants to continue pursuing research for longer than they would have if they had not received loan repayment? Without debt relief from the LRPs, participants presumably would take substantially longer to pay off their education loans; therefore, a substantial reduction in debt after two to four years in the program may relieve financial pressures that could otherwise influence participants to leave qualified research positions for higher-paying positions. But even after education debts are paid down in their entirety or in large measure, researchers presumably still face salaries lower than might be available to them in nonqualified research positions--and those higher salaries could be tempting.

[8] This is a slightly different question than might be asked when studying the recruitment effects of other programs. Other studies might ask how the *offer of program participation* affects job entry decisions. Such a question has natural appeal because it concerns an intervention that precedes the desired outcome. If that were the case here, we could compare the types of jobs taken by scientists offered loan repayment with those taken by a similar group of scientists not offered loan repayment and then attribute any differences to the offer. But that is not the case under the LRP operating rules. In fact, the opposite is true: LRP applicants must have arranged for a qualified position at the time of application, which precedes the award decision by several months.

In the end, participants' decisions to remain in or leave qualified research positions at any point probably reflect their life situations and the relative value or importance they attach to a host of pecuniary and nonpecuniary factors associated with the options available to them. However, it is difficult to predict which factors will have what level of importance to participants in diverse circumstances with diverse goals, needs, and so on. Accordingly, the evaluation should collect a good deal of information on study subjects' personal and professional circumstances and goals in order to permit analyses that will shed light on how retention effects may or may not be realized.

With respect to other potential effects on participants' careers, the major conceptual question is, How might acceptance into the LRP "cause" individuals to achieve various outcomes as researchers—for example, receive new NIH funding as a principal investigator, publish research articles, or attain tenure—that they would not have achieved if they had not received loan repayment? On the one hand, loan repayment might have only a limited effect on research productivity; loan repayment would probably not change participants' work responsibilities, afford greater opportunities, improve knowledge or skills, or increase ambition. On the other hand, the extra income could allow scientists who split their time between a highly lucrative clinical practice and less lucrative research activities to reduce the time spent in their clinical practice and devote more time to research. In the long run, debt reduction might reduce anxiety and allow program participants to be more relaxed or content, thus enabling them to concentrate more fully on their research and relieve them of the pressure of searching for a different (better-paying) job. Thus, by permitting individuals to devote more focused time to research, the LRPs might enable individuals to publish more articles, write better grant applications, and generally develop into more successful scientists than otherwise would have been the case.

III. RESEARCH QUESTIONS

Given the programs' stated goals, the two most important questions that an evaluation should ideally seek to answer are:

- Do the LRPs have a "recruitment effect" in that they increase the number of individuals who begin research careers in the designated LRP field?
- Do the LRPs have a "retention effect" in that they increase the length of time that individuals conduct research in the designated LRP field?

For reasons explained below, it may be possible to use only retrospective self-reports by applicants to gauge whether the programs affected recruitment. However, our proposed design rigorously addresses the second, and arguably more important, question. While a study could be limited to addressing that question alone, we believe it will be more informative to the NIH to expand the scope of the study to consider other potential impacts as reflected in the following two research questions:

- Do the LRPs increase the length of time that individuals conduct research in any biomedical field that falls under the mission of the NIH?

- Do the LRPs have a "productivity effect" in that Extramural LRP award recipients are more successful researchers than they would have been without the programs?

Later sections of this report address how the outcome measures could be defined and measured.

IV. CHOOSING A COMPARISON GROUP

To answer the research questions, we need to compare the outcomes associated with the Extramural LRPs to counterfactual outcomes—that is, what actually happened to participants versus what would have happened to them in the absence of the programs. Given that we cannot observe counterfactual outcomes directly, we must estimate them indirectly by, for example, measuring outcomes of a comparison group that is similar to program participants in every way except that its members did not receive Extramural LRP funding. One important decision is whether to select the comparison group from nonfunded members of the applicant pool (an internal comparison group) or from outside the applicant pool (an external comparison group). As explained in more detail below, a design that uses an internal comparison group would be feasible and allow for a rigorous evaluation of the effects of the Extramural LRPs on research retention and research productivity; in contrast, it does not appear to be feasible to construct a design that uses an external comparison group. While use of an internal comparison group will not allow a rigorous evaluation of recruitment effects, it will provide OLRS with information to gauge whether (and, if so, how) the programs might affect recruitment. Therefore, our recommendation is to use an internal comparison group.

A. Why Using External Comparison Groups Is Not Feasible

The primary difficulty in defining external comparison groups is that Extramural LRP recipients come from a wide variety of backgrounds. In particular, data supplied by OLRS indicate that recipients come from many fields (including medicine or fields such as biology, chemistry, health sciences, psychology, clinical psychology, and other social sciences), received their degree between 1 and 15 years before applying, work on research projects funded by a broad array of sources (including the NIH, other government agencies, and a wide range of foundations), occupy a variety of positions (including assistant professors, research fellows, interns, psychologists, and so forth), and work in many employment settings (including hospitals, universities, and research centers). In view of the diversity of recipients, the comparison group would also need to be broadly defined. While we attempted to develop several comparison groups that might allow us to measure LRP recruitment or retention effects, the need to use a broadly defined comparison group proved to be problematic.

1. Why Using External Comparison Groups to Measure Recruitment Is Not Feasible

One way to measure whether the LRPs had a recruitment effect would be to compare the rate at which two groups began pursuing research: those who were likely to be interested in and eligible for the programs and those who were "barely ineligible." In this case, program eligibles would be individuals who held a doctoral degree, carried debt equal to or greater than 20 percent

of their salary, and met all of a program's other requirements, whereas "barely ineligibles" would include individuals who would have met all of a program's eligibility requirements except that their debt totaled slightly less than 20 percent of their salary. We considered implementing this approach by using two comparison groups: MD degree holders and PhD degree holders.

Using MD degree holders as a comparison group turned out to be infeasible because the portion of MDs who would likely be conducting research in a particular LRP field is a small fraction of the pool of MDs**.** In particular, only 2 percent of MDs indicate that they are pursuing any type of research as their major professional activity. The percentage pursuing research in a particular field is even smaller. For example, it has been estimated that about a third of physicians indicating that research is their major professional activity are involved in clinical research (Committee on National Needs for Biomedical and Behavioral Scientists Education and Career Studies Unit, 2000). Therefore, the size of the sample required to measure whether the programs had a recruitment effect is prohibitively large. Even if data were collected on every physician in the United States (itself a daunting task), it would not be possible to detect even large effects of the Extramural LRPs. For example, even if the LRPs increased the number of physicians conducting clinical research from 5,000 to 5,500[9] or from 0.67 percent of the population of physicians to 0.73 percent, the effect would not be statistically significant at the 10 percent level.

Using PhD degree holders as a comparison group is infeasible because samples in available data sources are too small. We considered the use of a merged file of the Survey of Earned Doctorates, which contains debt information at the time of receipt of the doctoral degree, and the Survey of Doctoral Recipients, which contains occupation and income data during the year of the survey. From the merged files, we could construct debt-to-salary ratios for those who received their degree in the most recent cohorts. The merged files contain about 400 individuals who received their doctoral degree within the two years preceding the surveys and received their PhD in a field that is likely related to clinical research, including clinical psychology, health sciences, and biology. Unfortunately, the sample would not be sufficiently large to detect even the maximum possible impact of the Extramural LRPs. We estimate that the maximum impact of the programs on the percentage of PhD holders conducting clinical research is about 5 percent (800 PhDs receiving clinical research LRP award recipients divided by 15,000 estimated PhDs in the United States currently conducting clinical research).[10]

In summary, it does not appear to be feasible to attempt to measure whether the Extramural LRPs have recruited individuals to pursue clinical research from the pool of MDs or the pool of

[9] An effect of this size would require that about half of all MD clinical Extramural LRP recipients would not have pursued clinical research absent the clinical Extramural LRPs. (Approximately 1,000 MDs received clinical Extramural LRP awards during the first three years of the program.)

[10] This estimate of the number of PhDs conducting clinical research is drawn from a report on the need for scientists (Committee on National Needs for Biomedical and Behavioral Scientists Education and Career Studies Unit, 2000).

PhDs in relevant fields. It would be even less feasible to measure such an effect for the smaller LRPs.

2. Why Using External Comparison Groups to Measure Retention Is Not Feasible

We also had to rule out the use of an external comparison group to measure research retention. Defining a pool that would match the diverse backgrounds and work experiences of LRP participants would be difficult. For example, we considered constructing a comparison group of those in NIH training centers; however, such a group would not allow for an assessment of the programs' effect on the many individuals who do not come from such a training environment. Because it is difficult to imagine that any external comparison group would be able to match closely the characteristics of both funded and nonfunded applicants, we instead assessed the feasibility of using an internal comparison group (nonfunded applicants).

B. Using Internal Comparison Groups Is More Promising

Nonfunded applicants represent a possible comparison group because all were interested in the Extramural LRPs and have characteristics similar to those who received funding. Potential concerns with using nonfunded applicants include (1) selection bias stemming from funded applicants with greater perceived research potential than nonfunded applicants; (2) whether sample sizes will be sufficiently large to measure program effects; and (3) the difficulty of measuring recruitment effects since all applicants must already have a job in the designated LRP field to be eligible. However, these concerns can largely be overcome.

Selection Bias. Selection bias is a concern if funded applicants tended to have better outcomes (for example, longer careers in research in the designated LRP field) than nonfunded applicants even if the Extramural LRPs did not exist. Indeed, Extramural LRP awardees receive funding in part because the application review committee believed that they demonstrated the potential to persist in a research career. If the selection process is known and can be measured, however, we can use statistical models to adjust for selection and obtain unbiased estimates of the programs' effect.

Our review of applicant data indicates that much of the information used in the selection process was carefully recorded. As described earlier, ICs convened panels to score applications, and the average score was recorded. (One exception occurred in the case of the applications for the Health Disparities LRP and the Clinical Research for Individuals from Disadvantaged Backgrounds LRP, which were often "triaged"; that is, noncompetitive applications were not sent to the panel and did not receive a priority score.) In addition, other information drawn from the applications (such as type of degree, date of degree, conferring institution, debt levels, salary levels) has been stored electronically. Thus, data are available to control for differences in applicant quality (as measured by the application score) as well as for other possible differences in characteristics between funded and nonfunded applicants.

We next investigated whether the scoring process would allow us to use a regression discontinuity design. A regression discontinuity design can control for the selection bias inherent in the application process by testing for a sharp increase in outcomes just before and just

after a score cut-off point (see Figure 1 for an illustration). Such a design can be used if (1) assignment to the treatment group (that is, those who receive funding) is entirely or largely based on a continuous measure (such as a score) and (2) the score cut-off point that separates those who receive the treatment from those who do not receive it is chosen independently of the assignment of the scores. Our review of the LRP selection process suggests that the process met these requirements. In particular, at the time they scored the applications, reviewers who assigned a score to each applicant did not know the cut-off score (or range) that would separate those who would receive funding from those who would not. The average score an applicant received (which ranged from 100 to 500) was a major factor in ranking applicants to determine which ones did and did not receive funding. Each year, about half of the ICs funded only those above a certain cut-off score (though the cut-off score tended to differ for each IC, and, for those ICs that scored applicants for different LRPs, the cut-off score was different for each LRP). The other ICs used a "fuzzy" cut-off: they funded nearly all of the best-scoring applicants and almost none of the lowest-scoring applicants but considered factors other than the score in selecting applicants near the cut-off point for funding.[11]

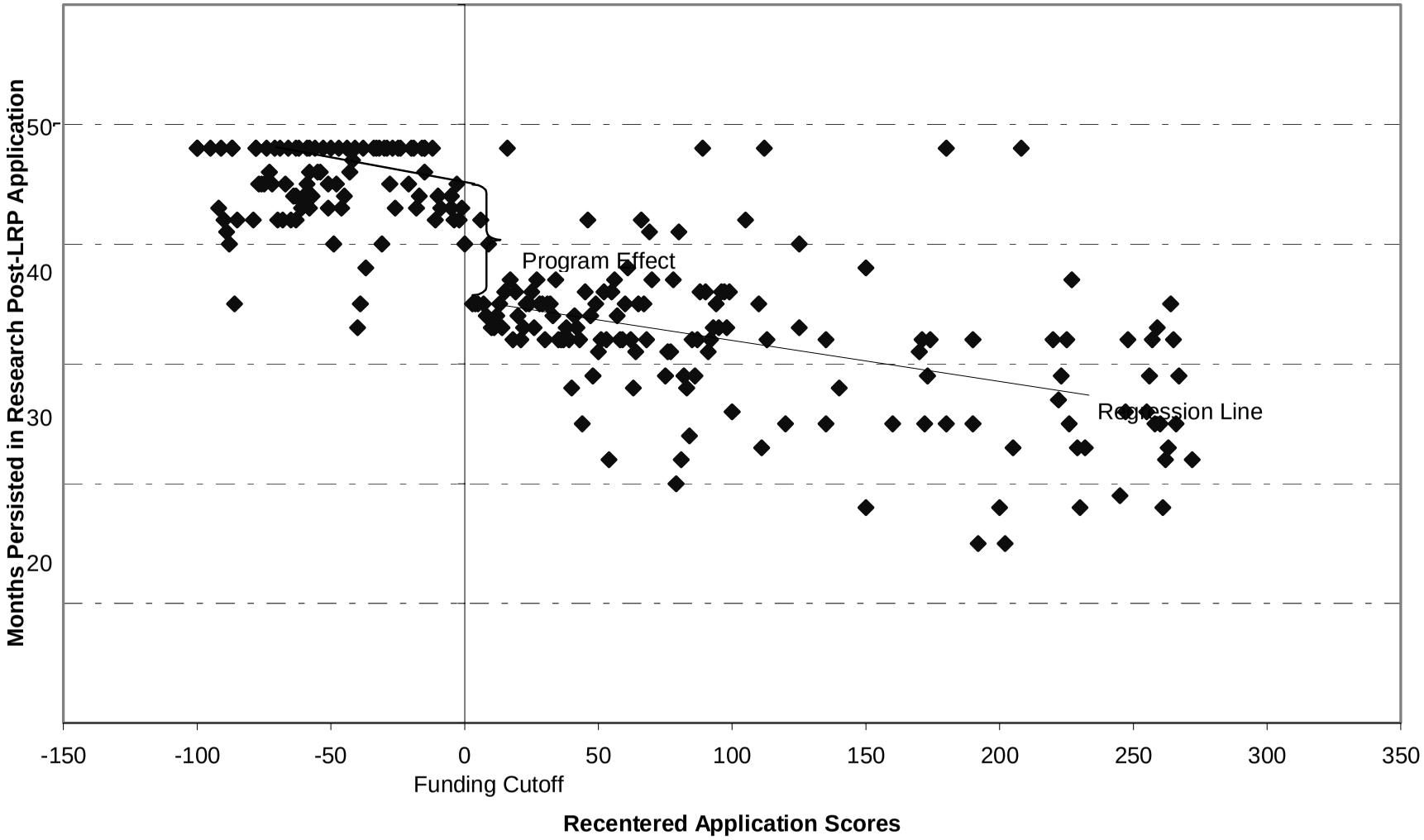
In short, it would be possible to measure rigorously the effect of the Extramural LRPs on research retention and research productivity by using nonfunded applicants as the comparison group, taking advantage of ways to control for the selection bias inherent in the application process. Section VI provides details about the methodology and how to control for selection.

Sample Sizes. A second concern associated with using nonfunded applicants as a comparison group is whether the sample size would be sufficient to detect program effects of a moderate size. Our analyses indicate that, for a sample including both the 2003 through 2004 cohorts, it would be possible to detect whether the Extramural LRPs collectively had an effect of 10 percentage points (see Section IV for additional details). This means, for example, that if the LRPs increased the percentage of individuals persisting in research careers from 50 to 60 percent, the effect would likely be statistically significant. We would also be able to detect effects of less than 15 percentage points for certain subgroups.

Measuring Recruitment Effects. A final possible concern associated with using nonfunded applicants for the comparison group is the difficulty of rigorously measuring recruitment effects for the simple reason that all applicants were already "recruited" in the sense that, at the time of application, they must have been funded in a job in a field relevant to one of the LRPs. However, a retrospective survey could be used to gauge useful information about whether the programs had a recruitment effect and how large that effect might be. For example,

[11] Our conversations with NIH officials were consistent with the data. Some officials said that they based their funding decisions strictly on the score. Others said that they ranked all applications by their score from best to worst but carefully examined all applications close to the "payline"; they did not necessarily fund applicants with, say, a score of 220 instead of those with a score of 240 if the applicant with the higher score was involved in high-priority research.

Figure 1: Effect of Extramural LRP on Length of Time Conducting Research (Hypothetical Example)



a survey could ask applicants (both funded and nonfunded) (1) whether they knew about particular Extramural LRPs before taking the research position they held at the time of application, (2) the extent to which they took the position because they knew they might have a chance of selection for an LRP award, and (3) how they gauged their chances of selection for a funding award.

If survey responses indicate that Extramural LRP applicants first learned about the programs after accepting a research position in an Extramural LRP field, then the LRPs cannot logically be recruiting individuals to certain fields of research. On the other hand, if responses indicate that applicants took a position in expectation of possible loan repayment, it is conceivable that the programs did have a recruitment effect. This finding would be strengthened if, among those who claim that they began research in a given field because of the Extramural LRPs, those not awarded funding left the field sooner than those who received funding.

In summary, the use of an internal comparison group is feasible and, given the nature of the selection process and the clear documentation of that process, would allow for a rigorous evaluation of the programs' effects on research retention. While reliance on an internal comparison group will not allow a rigorous evaluation of recruitment effects, it will provide OLRS with information about whether (and, if so, how) the programs might affect recruitment.

V. KEY VARIABLES AND DATA SOURCES

The key variables for the evaluation include outcomes (retention in the LRP field, retention in any research field, and indicators of research productivity) and individual-level characteristics that might be used as control variables in the regression analysis or to define subgroups. The proposed outcome variables follow:

- Whether still conducting research in relevant LRP field (the field of LRP to which the scientist originally applied)
- Whether still conducting research in any field
- Length of time conducting research in LRP field and in any field
- Percentage of time devoted to research in LRP field or other field
- Whether principal investigator on NIH grant
- Whether principal investigator on any grant
- Whether has NIH research funding
- Whether has any research funding
- Whether applied for NIH research funding
- Whether applied for any research funding
- Whether has a tenured academic position

- Whether conducting research in nonprofit or government setting
- Whether reviewer for the NIH or peer-reviewed journal
- Number of publications
- Number of citations to publications

The following are possible control or subgroup variables:

- Demographic characteristics (such as age, race/ethnicity, gender, marital status, and number of children)
- Educational background (including type and date of doctoral degree) and work experience
- Whether had NIH funding before applying
- Whether had been principal investigator on NIH grant before applying
- Level of qualified education debt as well as other debt
- Salary at time of application

Data sources will include (1) data that OLRS has collected from applications, (2) publications databases such as PubMed, (3) funding databases, including NIH's Information for Management, Planning, Analysis, and Coordination systems (IMPAC II) database, which tracks NIH awards and principal investigators, and the CRISP database (Computer Retrieval of Information on Scientific Projects), a searchable database of projects funded by the NIH and other government agencies, and (4) a proposed survey of past Extramural LRP applicants. Table 2 shows the type of data we intend to collect from these various sources.

We propose a survey because some information critical to the assessment of the Extramural LRPs will not be readily available from secondary sources. First, according to applicant data from OLRS, individuals have engaged in research funded by such a wide array of non–NIH sources that it would not be feasible to determine accurately from secondary sources whether non–NIH funding supported an individual; too many potential sources would have to be checked.[12] Second, while publications databases would permit the measurement of some of the

[12] Because the funding sources were recorded in OLRS application data in "free form" text fields, we did not attempt to estimate the number of sources that awarded funding to the applicant pool. However, this list of 10 randomly drawn funding sources provides a flavor of the variety of sources: Lustgarten Foundation for Pancreatic Cancer Research, Valley Mental Health, Arnold Leonard Cancer Research Fund, American Academy of Neurology Foundation, Sidney Kimmel Foundation, American Heart Association, Donald W. Reynolds Foundation, Juvenile Diabetes Foundation, American Cancer Foundation, and National Kidney Foundation.

TABLE 2

DATA SOURCES FOR KEY VARIABLES

	Applicant Data	Publications Databases (e.g., PubMed)	Research Funding Databases (e.g., CRISP, IMPAC II)	Proposed Survey of Applicants
Type of degree	X			
Years since doctoral degree	X			
Funding received before application	X			
Demographic characteristics	X			
Application score	X			
Whether currently conducting research in LRP field				X
Length of time pursuing research in LRP field				X
Whether conducting research in any field				X
Number of publications		X		X
Whether received NIH funding			X	X
Whether principal investigator on NIH grant			X	X
Whether principal investigator on Center for Disease Control and Prevention, Food and Drug Administration, Health Resources and Services Administration , or Agency for Health Care Research and Quality grant			X	
Whether an NIH reviewer			X	X
Whether referee for peer-reviewed journal				X
Whether NIH R-O1 grant recipient			X	X
Whether principal investigator on any grant				X
Whether received any research funding				X
Whether conducting research in any setting				X
Type of research setting (e.g., nonprofit, for-profit, government)				X
Whether in a tenure-track position				X
Whether LRP affected recruitment to research career				X

desired outcome variables, searching by name through publications databases will overstate the accomplishments of some (for example, by producing false matches for common names) and fail to record the accomplishments of others (for example, by failing to locate the publications of those who changed their name through marriage). Moreover, there are time delays between when research is conducted and publications are produced; a survey would allow us to capture information on research articles that have not yet appeared in print. Third, even though the NIH IMPAC II database and the CRISP databases track principal investigators (PIs), we do not know of databases with names of people who are conducting research as part of a team rather than as a PI. To our knowledge, no available database tracks those conducting research funded by the for-profit sector. Finally, a determination of whether an individual is conducting research relevant to a particular LRP, for example Health Disparities, would likely be best measured as a self-reported variable rather than by relying on evaluators to guess at the type or area of research.

While secondary data will be somewhat inaccurate (failing to record the accomplishments of some and overstating the accomplishments of others, as noted), the data are nonetheless an unbiased source of information; that is, we would expect such inaccuracies to occur equally for both funded and unfunded applicants. In contrast, survey data can be a biased source of information owing to a risk that funded applicants may be more willing to respond to a survey than unfunded applicants. Therefore, even though secondary data have limitations, we propose to use such data in combination with survey data. Specifically, by using both survey data and secondary data available in publications databases and the NIH database, we will be able to test the sensitivity of program impacts to survey nonresponse by comparing results based on the sample of survey respondents to results based on secondary sources. We also recommend the use of secondary data to measure some outcomes (such as whether an individual had NIH funding) because outcomes based on secondary sources will have larger sample sizes than those based on survey data (due to survey nonresponse).

The proposed survey of Extramural LRP applicants would ask individuals about their career path, their decision to apply for loan repayment, and their current involvement in research (funding received, publications, and so forth). We recommend a Web-based survey with telephone follow-up. Locating applicants should not be problematic as the applicant database should have fairly recent contact information for these individuals. Internet search engines should make it relatively easy to contact those who have relocated since the time of application, especially given that the sample is a high-profile population consisting mainly of physicians and academicians. On the other hand, experience from MPR's evaluation of the Intramural LRPs suggests that motivating individuals to respond to the survey can be difficult; they are busy professionals, and, in some cases, gatekeepers in the workplace may restrict telephone access to them. We expect to be able to draw on our experience in implementing the survey for the evaluation of the Intramural LRPs to suggest cost-effective ways to increase response rates. We will develop details on how the survey will be administered if OLRS agrees with our proposal to conduct a survey by exercising the contract option that would fund the development of a survey and an OMB package.

VI. METHODS FOR ANALYSIS

We recommend the use of a regression discontinuity design to analyze the data. The design tests for an increase in outcomes above and below the application cut-off score. In addition, the design could obtain unbiased estimates of the effect of the Extramural LRPs.[13]

There are two types of regression discontinuity designs: "sharp" and "fuzzy." A sharp regression discontinuity design requires funding decisions based strictly on application scores, where all those above the cut-off score receive the intervention (in this case, funding) and all those below the score do not. Recognizing that ICs do not always use a strict cut-off score, we propose a fuzzy regression discontinuity design[14] to test whether outcome measures rise sharply before and after the score cut-off range (that is, when the probability of receiving funding changes sharply). It allows for funding decisions to be influenced, but not completely determined by, application scores; factors other than the application score are used to make funding decisions within a range of application scores.

To implement the fuzzy approach, we would use a two-stage statistical model. The first stage predicts whether an individual would receive funding based on his/her application score, the funding cut-off or cut-off range of the IC that scored the individual, and other characteristics that might have been used in funding decisions but not captured in the score. For ease of interpretation, we would center the application score so that each IC's funding cut-off (or midpoint of cut-off range) is equal to zero. The second stage then tests whether the predicted funding measure is related to an individual's outcome above and beyond what would have been expected according to the application review committees' judgment of his/her promise as a researcher (as reflected in the application score).[15] See Appendix A for details. By controlling

[13] We also considered an alternative approach that would compare applicants with similar application scores, some of whom received funding and some of whom did not. The design would rely on the fact that there is considerable overlap in the scores of funded and nonfunded applicants, thus reflecting different IC funding thresholds. However, given that scores are not necessarily consistent across ICs, this approach would require us to convene a panel to rescore a sample of at least 300 to 500 applications to ensure that scores were consistent across ICs. While such an approach is advantageous in that it requires a smaller sample size than the regression discontinuity design, we assumed that convening a panel to rescore applications would not be acceptable to NIH decision makers.

[14] The fuzzy design model can be used both for ICs that used a fuzzy cut-off and those that used a sharp cut-off. As explained in the appendix, for ICs with a sharp cut-off, no scores fall into the fuzzy range.

[15] Several recent studies have used instrumental variables within a fuzzy regression discontinuity design (see, for example, Angrist and Lavy 1999 and Jacob and Lefgren 2004). The proposed design is most similar to that of Jacob and Lefgren (2004), who use standardized test scores as instrumental variables to predict which students were required to attend summer school and face retention; in their second-stage model, they used predicted summer school attendance and predicted retention to estimate the effect of extra time in school on student achievement.

for the application score in the model and then testing for a sharp increase in outcomes above and below the funding cut-off score, the two-stage model adjusts for the likely greater research promise of funded applicants as compared with nonfunded applicants.

One requirement of the regression discontinuity design is that it must properly characterize the relationship between the outcome variable and the variables and the model. Therefore, we would test variations of the model that allow, for example, for a nonlinear relationship between the application score and outcomes.

Our primary analysis would estimate the combined effect of all of the Extramural LRPs. In addition, we could run the model separately for subgroups such as for the larger LRPs (Clinical, Pediatric, and Health Disparities),[16] for those with NIH funding before application versus those who did not have funding, for those who received doctoral degrees recently versus longer ago, for those who had higher versus lower debt levels, and for those who received MDs versus PhDs. Subgroup analyses would test whether the programs have differential effects on people with different backgrounds or characteristics. For example, program effects might be greater for MDs, who presumably have more lucrative nonresearch career options, than for PhDs. Subgroup analyses might also guide OLRS in considering potential changes to eligibility requirements. For example, subgroup analysis would provide some insight into whether the programs were more effective for those who did or did not have NIH funding before applying for the LRPs. Likewise, subgroup analyses would provide OLRS with useful information if it wanted to consider refining rules pertaining to the type of degree that applicants must hold or how recently they need to have obtained their degree.

VII. OPTIONS FOR TIMING OF DATA COLLECTION AND SAMPLE SELECTION

A. Timing of Data Collection

The evaluation must strike a balance between providing information on the full impact of the programs on research careers, which may take many years to unfold, and providing information to policymakers in a timely fashion. The NIH officials we interviewed believed that many outcomes would not be measurable for five to seven years after an individual applied to a program, suggesting a lag until at least 2008 to 2010 for data collection on the 2003 cohort. However, by using a survey rather than secondary sources, we could measure some outcomes sooner. For example, a survey could collect data on short-term research retention and forthcoming publications before the publications are available in print. Another consideration in the timing of data collection is that locating individuals for a survey would likely be easiest if conducted close to the time when contact information was last collected (that is, when nonfunded applicants applied to a program and when funded applicants completed their contracts). Therefore, we recommend measuring *early* outcomes by collecting data four to five years from the time that individuals applied to a program; we also suggest measuring *long-term* outcomes by

[16] While we could also run separate models for the Contraception and Infertility and Clinical Research for Individuals from Disadvantaged Backgrounds LRPs, these programs are so small that is extremely unlikely that we would find any statistically significant program effects.

collecting data from secondary sources and/or from another survey seven to nine years after individuals applied to a program. (The technology and availability of searchable, linked electronic databases that could track research grants and publications is continually improving. By 2010, it might be easier than it is today to track an individual's research career through secondary sources, making a survey unnecessary.)

B. Sample Selection

When conducting the study, we would find it most efficient to draw a sample rather than attempting to collect data on the entire universe of individuals who ever applied to one of the Extramural LRPs. A primary consideration for the evaluation design is how large a sample is needed and how to select that sample from among the four cohorts of applicants to date. Given the small number of nonfunded applicants in the 2001 and 2002 cohorts,[17] we recommend that the sample include only applicants from the 2003 and/or 2004 cohorts. The larger the sample, the greater is the chance of detecting overall program impacts. A large sample is also particularly important if precise measurements of program impacts are desired separately for each LRP or for other subgroups, such as for MD degree holders versus PhD degree holders. Below we present options that would include (1) a census of all applicants in the 2003 and 2004 cohorts, (2) a census of all applicants in the 2003 cohort only, or (3) a census of all applicants in the 2003 and 2004 cohorts to the Clinical LRP only. The final choice of sampling option depends on OLRS's goals for the evaluation.

To construct the options, we first compiled tables to show the number of new applicants in each LRP (see Table 3)[18] as well as the number of applicants likely to be included in our sample (see Table 4). In the top panel in Table 4, we reduce the number of nonfunded applicants to avoid double counting those individuals not funded in one cohort who then reapplied for a program in a later cohort.[19] We also exclude those who did not receive a priority score because their applications were "triaged" (deemed noncompetitive), as such applicants would be excluded from the analysis. The top panel shows the likely sample sizes for outcome measures that can be captured through secondary data sources (for example, whether an applicant became

[17] In 2001, fewer than 50 extramural awards were made, and nearly all applicants were awarded funding. In 2002, 589 applicants received funding, but only 207 did not, and only 100 of those without funding had application scores.

[18] Renewal applicants would be included with the cohort in which they originally received funding.

[19] From our analysis of applicant data, we estimate that about a quarter of applicants who do not receive funding reapply in a subsequent year. We recommend classifying reapplicants with the most recent cohort in which they applied. By using everyone's most recent application score and funding decision, we assume capture of each individual's best (and most accurate) score. (Those who remain eligible and believe that their application could be significantly improved likely apply again; those who believe their application would be rejected again likely do not reapply.)

TABLE 3

NUMBER OF NEW LRP APPLICATIONS AND AWARDS

	2004		2003		2002	
Loan Repayment Program	Funded	Not Funded	Funded	Not Funded	Funded	Not Funded
Clinical Research	622	566	732	418	393	94
Clinical Research for Individuals from Disadvantaged Backgrounds	19	16	21	8	19	16
Contraception and Infertility Research	16	24	10	2	9	3
Health Disparities Research	132	139	106	57	112	58
Pediatric Research	228	276	298	196	168	36
Total	1,017	1,021	1,167	681	701	207

a principal investigator on an NIH grant). The bottom panel shows the likely sample size for outcome measures that would be collected through a survey (for example, whether the individual was still conducting research in the designated LRP field), assuming a survey response rate of 75 percent.

The smallest difference between outcomes for participants and nonparticipants that we are likely—with confidence—to be able to attribute to the programs, as opposed to attributing to other differences between the groups, is called the minimum detectable effect. Given certain sample sizes, we estimated the minimum detectable effects that we would obtain by using a regression discontinuity design, assuming a desired statistical power of 80 percent (see Table 5). In the case of a difference between funded and unfunded applicants of the size indicated in the table, the likelihood that we will detect a difference at conventional statistical significance levels (that is, at the 10 percent significance level for a two-tailed test) is at least 80 percent. A minimum detectable effect of 10 percentage points means that if the program increased, say, the percentage of individuals conducting research five years after applying to the program from 50 to 59 percent, that effect would likely be statistically insignificant; however, if the programs increased the percentage of individuals conducting research from 50 to 61 percent, that effect likely would be statistically significant.

To make these calculations, we had to make assumptions about what the mean of the outcome variables might be. We provide several scenarios because some outcome variables (such as whether an applicant became a principal investigator on an NIH grant) are likely to have much lower means than other outcome variables (such as whether an applicant is still conducting any research at all). For all of the scenarios, we assumed that the correlation between the

TABLE 4

NUMBER OF APPLICANTS AVAILABLE FOR SAMPLE

	2004			2003			2002		
	Funded	Not Funded, Excluding Estimated Number Reapplying in Later Cohort	Total	Funded	Not Funded, Excluding Estimated Number Reapplying in Later Cohort	Total	Funded	Not Funded, Excluding Estimated Number Reapplying in Later Cohort	Total
Panel A: Estimated Number of Individuals Who Had Application Priority Scores									
Clinical Research	622	425	1,047	732	314	1,046	393	71	464
Clinical Research for Individuals from Disadvantaged Backgrounds	19	2	21	21	1	22	19	0	19
Contraception and Infertility Research	16	18	34	10	2	12	9	2	11
Health Disparities Research	132	25	157	106	28	134	28*	0	0
Pediatric Research	228	207	435	298	147	445	168	27	195
Total	1,017	677	1,694	1,167	492	1,659	589	100	689
Panel B: Estimated Number of Survey Respondents Who Had Application Priority Scores									
Clinical Research	467	318	785	549	235	784	295	53	348
Clinical Research for Individuals from Disadvantaged Backgrounds	14	2	16	16	1	17	14	0	14
Contraception and Infertility Research	12	14	26	8	2	9	7	2	8
Health Disparities Research	99	19	118	80	21	101	na	na	na
Pediatric Research	171	155	326	224	110	334	126	20	146
Total	763	507	1,270	875	369	1,244	442	75	517

NOTES: (1) We assume that 25 percent of rejected applicants reapply in a subsequent year. Renewal applicants counted in the cohort in which they originally received funding. (2) A large fraction of nonfunded applicants in the Health Disparities and Clinical Research for Individuals from Disadvantaged Backgrounds LRPs did not receive application scores due to "triaging" (removal of noncompetitive applications from the panel review process). (3) Survey response rate assumed to be 75 percent.

*Because no nonfunded applicants were scored, there would be no comparison group; therefore, funded applicants are excluded from the totals for this LRP in 2002.

TABLE 5

MINIMUM DETECTABLE EFFECTS FOR SCENARIOS UNDER
REGRESSION DISCONTINUITY DESIGN

	Option 1 2003 and 2004 Cohorts		Option 2 2003 Cohort		Option 3 Clinical LRP, 2003 and 2004 Cohorts	
	Full Sample (n=3,353)	Survey Sample (n=2,515)	Full Sample (n=1,659)	Survey Sample (n=1,244)	Full Sample (n=2,093)	Survey Sample (n=1,570)
Minimum Detectable Effects (Percentages) for Binary Outcome Variable with Mean of .5						
Clinical Research	9.5	10.9	13.9	16.1	9.5	10.9
Clinical Research for Individuals from Disadvantaged Backgrounds	*59.9*	*69.2*	*211.5*	*244.3*	na	na
Contraception and Infertility Research	*117.9*	*136.1*	*160.1*	*184.8*	na	na
Health Disparities Research	26.8	31.0	43.9	50.7	na	na
Pediatric Research	18.6	21.5	20.8	24.1	na	na
Subgroup of 50 Percent of the Sample	11.2	13.0	15.7	18.1	13.4	15.4
All LRPs	7.9	9.2	11.1	12.8	na	na
Minimum Detectable Effects (Percentages) for Binary Outcome Variable with Mean of .2						
Clinical Research	7.6	8.7	11.2	12.9	7.6	8.7
Clinical Research for Individuals from Disadvantaged Backgrounds	48.0	55.4	*169.2*	*195.4*	na	na
Contraception and Infertility Research	*94.3*	*108.9*	*128.1*	*147.9*	na	na
Health Disparities Research	21.5	24.8	35.1	40.6	na	na
Pediatric Research	14.9	17.2	16.7	19.2	na	na
Subgroup of 50 Percent of the Sample	9.0	10.4	12.6	14.5	10.7	12.3
Pooled LRPs	7.6	8.7	8.9	10.3	na	na
Minimum Detectable Effects (Percentages) for Binary Outcome Variable with Mean of .7						
Clinical Research	8.7	10.0	12.8	14.8	8.7	10.0
Clinical Research for Individuals from Disadvantaged Backgrounds	*54.9*	*63.4*	*193.9*	*223.9*	na	na
Contraception and Infertility Research	*108.1*	*124.8*	*146.7*	*169.4*	na	na
Health Disparities Research	24.6	28.4	40.2	46.5	na	na
Pediatric Research	17.1	19.7	19.1	22.0	na	na
Subgroup of 50 Percent of the Sample	10.3	11.9	14.4	16.6	12.3	14.1
Pooled LRPs	7.3	8.4	10.2	11.8	na	na

NOTES: (1) Italics indicate effects that would be impossible to detect. (2) All tests assume 80 percent power, two-tailed test at 10 percent significance level (or one-tailed test at 5 percent significance level). (3) Correlation between whether individual received funding and application score assumed to be 0.8.

application score and the treatment variable (in this case, whether an individual received funding) was 0.8 (based on calculations of applicant data).[20]

All three design options discussed below assume that short-term outcome data would be collected from both a survey and secondary sources in 2007 or 2008, four to five years after individuals applied to a program. Longer-term outcomes could be collected seven to eight years after individuals applied to a program as part of a second, or follow-up, study. Data analysis and reporting of outcomes would occur in the year after data were collected. Table 6 presents proposed timelines.

1. Option 1 (2003 and 2004 Cohorts)

Overview and Tradeoffs. Option 1 would include in the sample all individuals from the 2003 and 2004 cohorts. In view of the large sample size (3,353 applicants; see Table 4), Option 1 would be able to detect the smallest program impacts, but it would be the most costly of the options because of the work involved in surveying such a large sample. Finally, the inclusion of the most recent applicant pool (2004) in the sample also means that data collection and analysis would occur later than in Option 2 (which does not include the 2004 cohort) because the 2004 cohort needs time to achieve various outcomes.

Minimum Detectable Effects. Under most scenarios, we could detect effects of 9 to 11 percentage points for the full survey sample, depending on the mean of the outcome variable. We would also be able to detect effects for the larger LRPs (Clinical and Pediatric) of 10 to 20 percentage points. For subgroups comprising half the sample, we could detect effects of 13 to 15 percentage points.

2. Option 2 (2003 Cohort Only)

Overview and Tradeoffs. Option 2 would include only the 2003 cohort. Compared with Option 1, it would be less costly, requiring a sample about half as large. Another advantage is that data collection could begin a year earlier than in Option 1 because the 2003 cohort would have one more year than the 2004 cohort to achieve various outcomes, yielding results sooner for OLRS.[21] For example, data for Option 2 could be collected on short-term outcome sources in 2007 (four years after individuals applied to a program), whereas the same data would be collected in 2008 for Options 1 and 3. However, under Option 2, the minimum detectable effects

[20] In a regression discontinuity design, the high correlation between the treatment variable and the application score means that larger sample sizes are required than those required in experimental designs. In particular, the denominator of the variance of the regression discontinuity estimator includes the term $(1-R^2)$, where R is the correlation between the application score and whether an individual received funding.

[21] Alternatively, data collection and analysis could begin a year later (in 2008) but provide five-year rather than four-year outcomes, if timeliness of the evaluation is less important to OLRS than measuring longer-term outcomes.

TABLE 6

KEY FEATURES AND TIMING OF IMPACT EVALUATION OPTIONS

	Option 1	Option 2	Option 3
Cohorts Included	2003 and 2004	2003	2003 and 2004
LRPs Included	All	All	Clinical
Estimated Applicant Sample Size	3,353	1,659	2,092
Estimated Survey Sample Size	2,515	1,244	1,569
Minimum Detectable Effects for Survey Respondents	About 9 percentage points	10 to 13 percentage points	9 to 11 percentage points
Minimum Detectable Effects for Subgroup of Half of the Sample	10 to 13 percentage points	15 to 18 percentage points	12 to 15 percentage points
Advantages	Ability to detect smallest program effects; will be able to detect moderately sized program effects separately for the larger LRPs	Relatively early analysis and reporting; least costly of options due to smaller sample	Middle of three options in terms of cost and size of minimum detectable effects
Disadvantages	Most costly due to large sample; timing of analysis and reporting later than Option 2	Minimum effects are greater than in other options	No results for other LRPs; timing of analysis and reporting is later than for Option 2
Design Survey	Summer 2005	Summer 2005	Summer 2005
Possible Pilot Survey/Pretest	Pretest survey in 2007	Pretest survey in early 2006	Pretest survey in 2007
Conduct Survey	2008	2007	2008
Collect Data from Secondary Sources	Early 2008	Early 2007	Early 2008
Analyze Data	Fall 2008	Fall 2007	Fall 2008
Draft Report	2009	2008	2009
Collect Long-Term Outcomes from Secondary Sources, Possible Survey	Early 2011	Early 2010	Early 2011
Analyze and Report on Long-Term Outcome Data	Fall 2011	Fall 2010	Fall 2011

would be larger than in Options 1 and 3, and our ability to measure impacts for subgroups would be reduced.

Minimum Detectable Effects. Under Option 2, we would be able to detect effects of 10 to 13 percentage points for survey respondents, depending on the mean of the outcome variable. For the larger LRPs, we could detect effects of 13 to 25 percentage points. For subgroups consisting of half of the sample, minimum detectable effects are about 15 to 18 percentage points.

3. Option 3 (Clinical LRP Only)

Overview and Tradeoffs. While Options 1 and 2 would pool data on applicants to all five Extramural LRPs, OLRS might want to begin by analyzing data for only one LRP. Such an evaluation could be conducted for the largest LRP (Clinical) and include the 2003 and 2004 cohorts. Option 3 requires a moderately sized sample (2,092 applicants, which is larger than Option 2 but smaller than Option 1). It would therefore be less expensive to implement than Option 1 but would still allow us to detect relatively small effects for the Clinical LRP and moderate effects for subgroups that split the sample in half. The main drawback is obviously that results would not be available for the other LRPs. (OLRS might choose to evaluate the other LRPs later, depending on the success of the evaluation of the Clinical LRP.)

Minimum Detectable Effects. Under most scenarios, we would be able to detect effects of 10 percentage points or less for the Clinical LRP. For subgroups that included about half of the sample, we would be able to detect effects of about 15 percentage points or less.

4. Recommendations

The choice of an option should depend on OLRS's priorities regarding what type of evidence it wants on Extramural LRP effectiveness and how soon it needs that evidence. Cost may also be a consideration. If OLRS desires separate estimates for the larger LRPs and for other subgroups, then we would recommend a design that includes all applicants from the 2003 and 2004 cohorts (Option 1). If subgroup estimates are not a priority, then we would recommend Option 2, which will produce solid overall estimates of program effectiveness at a relatively lower cost because of a smaller sample size. If it is most important for the evaluation to produce impact estimates as soon as possible, then Option 2 would also be the best choice. One attractive option might be to include in the sample only applicants from the Clinical LRP; we could detect fairly small effects by using a smaller sample than in Option 1, although results would not be available for the other LRPs.

In assessing which option to pursue, OLRS should consider that, in MPR's experience in program evaluation, it is rare to see impacts much greater than 10 percentage points. The smallest minimum detectable effects that could be detected in any of our proposed options range from 7 to 10 percentage points. If OLRS would like to detect smaller impacts, the evaluation would need to be delayed until the pool of individuals who have ever applied to and participated in the LRPs has increased. On the other hand, OLRS may decide that effects smaller than 7 percentage points are not large enough to deem the LRPs cost-effective. That is, OLRS might believe that the cost of the Extramural LRPs (over $60 million per year) is justified only if the programs increase the percentage of scientists retained in a particular field by more than 7 percentage points; if such is the case, it is not a problem if the evaluation fails to detect smaller program effects.

VIII. PRODUCT OF EVALUATION

The product of the evaluation would be a report providing estimates of the collective impact of the Extramural LRPs. The analysis would control for applicants' characteristics such as the LRP to which they applied, type of degree, and debt levels. For some options, the report would

likely be able to provide separate estimates of the impacts of each of the two largest LRPs (Clinical and Pediatric) and possibly of the Health Disparities LRP. Given the size of the two smallest programs, it would not be possible to estimate their impacts with any precision; however, the average outcomes for participants in these LRPs would be reported. The report would also provide estimates for subgroups such as MDs and PhDs. Table 7 is an example of how the report would present the results.

TABLE 7

SAMPLE TABLE ON THE EFFECT OF EXTRAMURAL LRPs ON OUTCOME MEASURE (FOR EXAMPLE, PERCENTAGE CONDUCTING RESEARCH IN LRP FIELD)

	Effect of Extramural LRPs	Significance Level
Clinical Research		
Health Disparities Research		
Pediatric Research		
All Five LRPs*		
Subgroups		
Holds MD		
Holds PhD		
Had NIH funding before applying		
Did not have NIH funding before applying		
Received most recent doctoral degree within past five years		
Received most recent doctoral degree more than five years ago		
Had debt in excess of $70,000		
Had debt less than $70,000		

*Not possible to generate separate estimates for the Clinical Research LRP for Individuals from Disadvantaged Backgrounds or the Contraception and Infertility Research LRP because there are too few applicants with scores.

IX. CONCLUSION

A design that uses nonfunded applicants as the comparison group is feasible and should be able to provide unbiased estimates of the effectiveness of the Extramural LRPs. We developed three options for which cohorts and LRPs should be included in the sample, although other options might be possible after we receive feedback from OLRS on its research priorities and preferences. The final choice of option depends on balancing cost considerations with the desire to measure impacts precisely for subgroups. After receiving feedback from OLRS, we will make a final recommendation about which option to implement and produce a final design document.

REFERENCES

Angrist, Joshua D., and Victor Lavy. "Using Maimonides Rule to Estimate the Effect of Class Size on Student Achievement." *Quarterly Journal of Economics,* 114(2) (1999), pp. 535-575.

Committee on National Needs for Biomedical and Behavioral Scientists Education and Career Studies Unit, Office of Scientific and Engineering Personnel, National Research Council. "Addressing the Nation's Changing Needs for Biomedical and Behavioral Scientists." Washington, DC: National Academy Press, 2000.

Jacob, Brian, and Lars Lefgren. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics* LXXXVI.1 (2004), pp. 226-244.

APPENDIX A

STATISTICAL MODEL

To estimate the impact of the Extramural LRPs, we propose the use of a fuzzy regression discontinuity design whose implementation requires a two-stage model. In the first stage, we predict whether an individual received funding by using instrumental variables indicators for whether an individual is above or below the scoring cut-off point (or range). In the second stage, we predict outcomes based on a predicted funding indicator (from the first-stage equation), the continuous version of the application score, and other variables.

In the first-stage model, we would estimate for individual i:

$$(1)\ PFUNDED_{i=} B_o + B_1RANGE_i + B_2CUTOFF_i + B_3*SCORE_i + B_4X_i + E_i$$

where $PFUNDED_i$ is the predicted probability that an individual received funding; B_o is an intercept; $CUTOFF_i$ is a binary variable indicating whether an individual's application score was better than the cut-off score or cut-off range used by his/her IC (the IC that scored the application); RANGE is an indicator equal to 1 if an individual's application score fell in the cut-off range for his/her IC (the indicator would be zero for an individual whose score was outside the cut-off range and zero for an individual scored by ICs that used strict cut-off scores); SCORE is the individual's recentered application score; X is a vector of individual background characteristics (such as field of research or type of degree) that might have been used in funding decisions but not reflected in the score; and E is an error term. We would expect the application score to be virtually unrelated to funding decisions that were outside the cut-off range (that is, the coefficient, B_3, would be close to zero). To the extent possible, the first-stage equation would be run separately within each LRP, IC, and cohort; ICs that scored only a few applications would be grouped with other ICs with similar score cut-offs for the purpose of the equation.

The second-stage equation for individual i would be:

$$(2)\ Y_i=B_o +B_1PFUNDED_i +B_2SCORE_i + B_3X_i + B_4IC_i + B_5COHORT_i +e_i$$

where Y is an outcome measure; B_o is the intercept; PFUNDED is the predicted probability that an individual received funding (from equation 1); SCORE is the application score; X is a vector of individual background characteristics such as type of degree, race, and age; IC represents a vector of IC dummy variables indicating the IC that scored the applicant; COHORT is a dummy variable capturing the year that the individual applied for funding; and e is an error term.[22] The effect of receiving Extramural LRP funding is captured by the coefficient on predicted funding, B_1.

[22] While some models would be run separately for each LRP, in models that pool the LRPs, we would also include dummy variables that represent each LRP.

The identifying assumption in the fuzzy regression discontinuity design is that the instruments (that is, the discontinuous measures of the score cut-off, CUTOFF and RANGE) are related to outcomes only through their effect on whether an individual received funding; we assume that our continuous version of the application score (SCORE) adequately controls for the relationship between application scores and outcomes. If equation (2) is correctly specified, we would falsely conclude there was a program effect only if some other factor caused an increase in outcomes right at the score cut-off point or range. It would be highly implausible for some other factor to cause an increase in outcome right at the cut-off point.

An underlying assumption of the regression discontinuity design is that equation 2 is specified correctly. In particular, the model specified in equation 2 assumes a linear relationship between the score and the outcome variable, which may not be correct. However, the model can be modified to test for a nonlinear relationship between the score and outcome variables by including in the model, for example, the application score squared or application score cubed. The model can also be modified to include interaction terms between whether the applicant received funding and the application score, if, for example, those with lower scores benefit more from the LRPs than those close to the cut-off. Sensitivity tests would indicate whether results are stable for different specifications of the model. If estimates of the effect of the Extramural LRPs change depending on which model is used, we would recommend erring on the side of overfitting the model (that is, including extra polynomial terms and interactions) because the inclusion of extra terms yields unbiased coefficients (but reduces statistical power).

APPENDIX B

SUMMARY OF INTERVIEWS WITH NIH STAFF

Introduction

As part of our work in developing a draft design for an evaluation of the Extramural LRPs, we conducted brief telephone interviews with officials from eight ICs to build our knowledge of program operations and inform our decisions regarding various design issues. Major topics of discussion included the application review and funding process, possible outcome measures and time frames for collecting information about the review and funding process, and potential data sources. We selected representatives of ICs that awarded a relatively large proportion of the new contracts funded in FY 2003, as shown below. In most cases, the respondents were the individuals listed as an IC's program liaison on the LRP Web site, but in some cases the liaison referred us to another knowledgeable individual.

Institute/Center	Rationale for Selection	Respondent(s)
National Cancer Institute (NCI)	Awarded the most (108 of 729) contracts for the Clinical LRP and 30 of the 300 Pediatric LRP contracts	Dr. Carolyn Strete
National Heart, Lung, and Blood Institute (NHLBI)	Awarded 83 of the 729 Clinical LRP contracts and 36 of the 300 Pediatric LRP contracts	Dr. Larry Friedman
National Institute of Allergy and Infectious Diseases (NIAID)	Awarded 65 of the 729 Clinical LRP contracts and 38 of the 300 Pediatric LRP contracts	Dr. Milton Hernandez
National Institute of Mental Health (NIMH)	Awarded 97 of the 729 Clinical LRP contracts	Dr. Mark Chavez
National Center on Minority Health and Health Disparities (NCMHD)	Awarded all 106 of the Health Disparities LRP contracts and all 21 of the Disadvantaged Background LRP contracts	Ms. Kenya McRae, Dr. Lorrita Watson
National Institute of Child Health and Human Development (NICHD)	Awarded all 10 of the Contraception and Infertility LRP contracts and 50 of the 300 Pediatric LRP contracts	Dr. Eugene Hayunga
National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)	Awarded 34 of the 300 Pediatric LRP contracts	Dr. Francisco Calvo
National Institute on Drug Abuse (NIDA)	Awarded 59 of the 729 Clinical LRP contracts	Dr. Teresa Levitin

We did not intend our sample to be statistically representative of all NIH stakeholders in similar positions; rather, our goal was simply to explore some major issues with officials who have a relatively high degree of familiarity with one or more of the LRPs. Below we summarize the information and opinions offered by the respondents.

Applicant Review and Funding Process

Respondents provided fairly consistent accounts of the application review process. Applications are sent to ICs based on the substantive type of research the applicant will be conducting. ICs convene panels of outside experts, and applications are divided among panelists, with each application assigned to a few primary reviewers. The primary reviewers independently score the applications on a scale of 1 to 5, with 1 the highest score. Reviewers do not concern themselves with "the science" or with applicants' debt levels but rather focus on applicants' potential and plans for successful research careers. Their ratings of a given application are reportedly very similar in most cases. They discuss their ratings for the benefit of all panel members, who then assign their own scores to the application. The scores are averaged and then multiplied by 100, and applicants within each IC are ranked from high to low, with separate lists for each LRP. IC staff then make decisions about who should or should not be funded.

At this point in the process, ICs may follow different approaches. Sometimes an IC will simply go down the list strictly in order and fund as many applicants as possible, given the available funding. One respondent guessed that it might be easier for ICs to follow a top-down approach when their available funds come close to covering the collective need of all applicants. Other times, however, an IC may pass over an applicant at some point in the distribution to offer funding to someone lower on the list. One of the important issues that some respondents were able to address was ICs' decisions to veer from a straight, top-down funding process. In some cases, an applicant with a very high score, well above the likely cut-off point, might be passed over. One respondent explained that IC officials may see the person as already so well established in a research career that the funding would be unlikely to influence his/her retention. In other cases, officials might make exceptions much closer to the cut-off point. Two respondents explained that once they have a sense of where the cut-off point will fall, they might look at applicants slightly below the line to see if any are involved in research in particularly important or high-priority areas; such individuals might be selected in lieu of others with slightly higher scores. One of these respondents also explained that a similar preference might go to a reapplicant looking for two additional years of funding.

The IC forwards to the Office of Loan Repayment and Scholarship their recommendations for which applicants should be funded. If OLRS determines that someone recommended for loan repayment is no longer eligible or interested in funding, a funding opportunity could open up for someone else not initially on the recommended list.

Outcome Measures

Recruitment Goal. NIH Web sites describe the Extramural LRPs as intended to foster the recruitment and retention of researchers in particular fields. Given that interested individuals must have a qualified position in order to be eligible for the programs—that is, they must enter the field before they may receive the award—we asked nearly all respondents about how they viewed recruitment as a program goal and/or the extent to which they thought the programs could achieve entry effects. A few confirmed that recruitment into particular fields is an apparent program goal, but only one believed that it was as important as retention. Most respondents viewed retention in research as a more important and feasible program goal and

thought it unlikely that (or questioned whether) people would take an eligible research position (over a noneligible one) based on the prospects of winning the award. Rather, they believed that people took eligible research jobs because they had appropriate training and a pre-existing interest. One respondent described loan repayment as "icing on the cake" for many participants.

Type of Research. Because the Extramural LRPs are aimed at promoting four specific types or areas of research—clinical, pediatric, health disparities, contraception and infertility—we asked respondents how important they feel it is, if at all, for participants to remain in the same fields after leaving the LRPs. For example, should someone funded by the Pediatric LRP be expected to continue conducting pediatric research, or would switching out of pediatric research be an acceptable outcome? We heard a full range of opinions. Three respondents said that it is an important goal for LRP participants to remain in the same field and that switching is definitely not a desired outcome.[23] Another respondent said that if participants changed their area or type of research, he would wonder whether the NIH had wasted its money; the same respondent nonetheless believed that performing some type of research would be better than performing none at all. Similarly, another respondent said that research in the LRP field is the preferable outcome but that research in another field was acceptable and preferable to no research. And four respondents felt that switching fields was an acceptable outcome and should not be considered a sign of program ineffectiveness; all four stressed, however, that participants should ideally remain involved in some type of research as opposed to none at all.

Portion of Time Devoted to Research. Given that the LRP eligibility rules require applicants to devote at least 50 percent of their time to research in the designated field, we asked some respondents whether that minimum standard, or any other, should be used to gauge whether individuals are sufficiently engaged in research. Those who addressed the issue either had no opinion or believed that the portion of time devoted to research was not particularly important as compared with other outcomes or other indications that someone is conducting research.

Employment Sector. Recognizing that the LRP eligibility rules require applicants' research to be funded by a domestic nonprofit entity (a foundation, professional association, other nonprofit institution, or a federal, state, or local government agency) and that applicants may not be full-time federal government employees, we asked some respondents whether employer type would be an important outcome to consider for former LRP participants. For example, if someone were engaged in research at a for-profit corporation, would that be a positive or negative outcome? Respondents were somewhat divided in their opinions. A few said that the setting in which former LRP participants are working would not be an important consideration, emphasizing again that "doing research is what matters," not the nature of the employer or funding source. One respondent felt that working in the private, for-profit sector would be an acceptable outcome, but not as optimal as working in the public or nonprofit sector.

[23] One respondent qualified this assessment by saying that if participants were not doing *good* research, not making a valuable contribution, then their departure from the field, or from research in general, would actually be a good outcome. The same respondent also suggested that, if possible, a study should query participants about why they left the LRP field or research in general to see what role financial considerations played.

And one respondent said that the spirit of the programs was to promote "academic-"style research; therefore, if a former LRP participant were working at a drug company, for example, the outcome would be deemed undesirable.

Research Administration. We asked several respondents how they would view former LRP participants working in research administration. One respondent saw research administration as a valuable outcome. Another respondent agreed but, along with another respondent, noted that most participants are too early in their careers to be taking such positions. One respondent would not worry about someone moving into research administration if the person eventually intended to go back into research. Finally, four respondents said that research administration would definitely be a less desirable outcome than conducting research.

Basic Research Indicators. With all respondents, we discussed possible indicators that someone is engaged in research—especially factors that could be detected through secondary data sources as opposed to simply asking people if they are conducting research. Suggested indicators included applying for grant funding, publishing research articles, and participating in professional meetings or conferences.

Degree of Success in Research. Although all respondents emphasized the importance of LRP participants continuing in research after leaving the LRPs, some suggested that a study of career outcomes among former LRP participants should consider individuals' degree of success in research. As one respondent said, "It's not the amount of [time in] in research, it's the *quality*." Respondents seemed interested in knowing the extent to which participants were living up to the potential reflected in their selection to receive loan repayment; in short, they would apparently expect participants to have more impressive careers, on average, than nonparticipants. Suggested indicators included publication productivity; publishing as a lead or senior author; extent to which publications are cited in other researchers' publications; types of grant funding received as a principal investigator (with NIH R-01 grants considered most prestigious, possibly followed by National Science Foundation grants and then grants from other foundations); receipt of honors or awards; type of academic appointment; and whether they are functioning as independent researchers, responsible for their own laboratory and pursuing their own research program, as opposed to serving on a research team but not as a PI. Other comments, however, reflected a view of outcomes that might not be as easily or objectively measured. The respondent quoted above, for example, wanted to know whether the LRPs enable participants to perform *better* research than they otherwise would and to have a more positive impact on public health. A final note on this subject: the respondents came up with these suggestions on their own; we did not ask questions about indicators of quality, prestige, or success.

Outcome Data Sources

We asked respondents for suggestions of data sources that might be useful for determining whether LRP participants had achieved some of the outcomes discussed earlier. Virtually all respondents mentioned research publications databases such as PubMed as a useful source for determining whether participants had any publications to their credit.

Suggested sources of information on grant funding included the NIH's IMPAC II (Information for Management, Planning, Analysis, and Coordination) and CRISP (Computer

Retrieval of Information on Scientific Projects) systems, National Science Foundation, Health Resources and Services Administration, Centers for Disease Control, Substance Abuse and Mental Health Services Administration, U.S. Department of Defense, U.S. Department of Veterans Affairs, U.S. Department of Energy, American Cancer Society, American Heart Association, National Kidney Foundation, Juvenile Diabetes Research Foundation, Cystic Fibrosis Foundation, Robert Wood Johnson Foundation, Howard Hughes Medical Institute, and Bill & Melinda Gates Foundation. One respondent pointed out that clinical researchers could be working in virtually any substantive area, making it difficult to single out particular data sources worthy of consideration. In naming foundations and other funding sources outside the NIH, most respondents were simply indicating an awareness of the organization's major role in funding research; they did not know the extent to which the types of data that we might need are maintained or could be obtained—for example, whether an organization maintains records of all key staff working on a project or just the principal investigators.

One respondent suggested that basic employment or occupational data might be obtained from large universities' alumni databases.

Measurement Timing

In discussions with all respondents, we asked how long they felt it would take for LRP participants to achieve the types of outcomes they saw as important indicators of program effectiveness. Without specific prompting, a near consensus emerged: an estimate of five years after loan repayment recipients stopped participating in the LRPs. Respondents explained that it can take a couple of years to secure a grant, two years for a publication to appear after the end of a project, and three years for a publication to be cited in another publication. One respondent pointed out that the appropriate interval between program completion and outcome measurement would differ with the participant's background or experience—for example, for someone receiving loan repayment while an assistant professor, it might be reasonable to consider outcomes achieved within two or three years, whereas for a more junior LRP recipient, it would be appropriate to wait at least five years. Two respondents noted that it might be reasonable but not ideal to measure outcomes two or three years after program participation; they warned that the picture could change substantially in the subsequent few years. One respondent thought it would be worthwhile to study participants even one year after their commitment ends in order to develop an early sense of the extent to which they leave or stay in their LRP fields.

Concluding Observations

Our interviews were useful in various ways. First, they helped us develop a deeper, clearer understanding of how the Extramural LRPs operate, particularly the application review process and funding decisions. Such insight is necessary for interpreting the award patterns observed across ICs. Second, comments about common career paths and research activities among biomedical researchers were useful in highlighting the potential advantages and disadvantages of measuring outcomes at different points after application (or program completion).

Third, opinions on basic issues such as desirable or expected program outcomes suggest that some stakeholders in the NIH are likely to have different interpretations of the potential results of an impact study.

- If, for example, results were to show that former program participants are no more likely than a comparison group to be involved in research in their LRP fields (clinical, pediatrics, health disparities, contraception and infertility) but are more likely to be participating in some type of research, some stakeholders would see such involvement as a positive outcome, thus indicating that the programs have a socially desirable effect. Other stakeholders would see it as a negative outcome, thus indicating that the programs are not achieving their desired effect.

- If, on the other hand, an impact study could be designed as a rigorous exploration of recruitment effects but then finds little or no impact, some stakeholders might downplay or dismiss the results, saying that it was unrealistic to expect a positive impact.

- Similarly, if a study were to find no differences between participants and a comparison group in terms of a relatively blunt outcome measure—such as simply whether individuals were "engaged in research" at a certain time—some stakeholders might argue that the more important findings pertain to degree of success in research.

While it is important that an impact study focus on a program's formal or official desired outcomes and define and measure them appropriately, it may also be useful to explore other outcomes, or alternative definitions of those outcomes, to ensure that the study has credibility with diverse stakeholders.

Finally, many respondents may have assumed that outcomes would be gauged exclusively through secondary data sources such as publication indices. If such an approach were taken, then a relatively long time frame might indeed be necessary—as the respondents pointed out, for example, it could take two years for a research article to appear in print. However, we would note that a more direct approach, such as survey, would address similar outcomes sooner. For example, a survey could ask individuals about the number of journal articles they submitted for publication, the number under consideration, and the number accepted and awaiting publication.