Feasibility Study for Evaluating Research Training in NIMH, NIDA, and NINDS

Final Report

Authors:

Susan T. Azrin, PhD
Howard H. Goldman, MD, PhD

Date:

May 4, 2005

Prepared for:

National Institute of Mental Health

National Institute on Drug Abuse

National Institute of Neurological Disorders and Stroke

Bethesda, Maryland

Prepared by:

WESTAT
Rockville, Maryland

Contents

1. Introduction and Background

OVERVIEW

Three National Institutes of Health (NIH) Institutes—the National Institute of Mental Health (NIMH), the National Institute on Drug Abuse (NIDA), and the National Institute of Neurological Disorders and Stroke (NINDS)—asked Westat to conduct a feasibility study for evaluating these Institutes' research training programs. The *feasibility* of an evaluation is defined as the extent to which an evaluation is appropriate and practical for implementation (Joint Committee on Standards for Educational Evaluation, 2003). For the current feasibility study, this meant determining if, for example:

- the desired results of NIMH, NIDA, and NINDS research training, across the Institutes' research training mechanisms, can be readily identified and agreed upon;
- the outcomes to measure these desired results can be agreed upon; and
- appropriate data to measure these outcomes is readily available and of sufficient quality, or can be collected without undue cost or burden.

Westat conducted this study over an 8-month period between September 2004 and April 2005. We conclude that the major stakeholders in the NIMH, NIDA and NINDS research training programs, particularly the three Institute Directors, would like a fuller picture than currently available on the career outcomes and scientific outputs of their research trainees, in addition to answering the key research questions posed in this feasibility study. Not surprisingly, the three Institutes are most interested in the outcomes for the research training awards to which the Institutes have devoted the most resources. Those awards are the pre- and post-doctoral T32, F31, F32, and mentored K awards, including the K01, K08, K12, K22, K23, and K25.

In this chapter we provide the background and motivation for the feasibility study. We then outline the feasibility study design and methods employed in *chapter 2,* and summarize the feasibility study findings in *chapter 3.* In *chapter 4* we present an evaluation design, including key research questions, specific outcomes to be measured and approaches to measuring them, and plans for sampling and data analysis.

NEED FOR AN EVALUATION

Of the NIH's fiscal year 2005 budget, 2.7 percent or $762 million is dedicated to research training. Despite this substantial investment in research training, the NIH has never before conducted a comprehensive, multi-Institute evaluation of research training outcomes. The NIH has reasons for wanting to carry out such an evaluation now.

In an era of slow NIH budget growth and limited resources, Institutes must set their priorities carefully in order to fulfill their scientific missions. Part of this priority setting involves striking the proper balance between investment in research training and investment in independent investigator-initiated research. Within the research training budget, an Institute must also make many choices as to where and how to deploy its research training resources. For example, at the pre- and post-doctoral levels of research training, an Institute may choose to emphasize investment in institutional research training programs by using National Research Service Award (NRSA) institutional training awards (T awards), versus investment directed to the individual through NRSA individual-level research training awards (F awards). At what stage of research training an Institute's investment yield the most benefit—at the pre-doctoral, post-doctoral, or early career scientist stage? Such decisions are difficult for an Institute to make in the absence of data on the results of its research training investments.

An evaluation of an Institute's research training investment could show to what extent this training contributes to the Institute's overall objectives and goals and produce information on how to improve both the Institute's research training programs and NIH research training programs in general. Consequently, three Institutes with related missions, NIMH, NIDA, and NINDS, would like to conduct a comprehensive evaluation of their research training programs. It is anticipated that the evaluation findings, as well as the methodology and tools developed in the course of conducting the evaluation, will also be of interest and use to other Institutes.

Training program evaluation at NIH previously has only been undertaken for specialized training programs, such as programs to increase diversity in the scientific workforce. Small-scale studies of certain elements of the NIH's research training programs have been examined, such as the early career progress of NRSA pre-doctoral trainees and fellows (Pion, 2001), a focus group study of early career clinical researchers in the K23 award program (Henderson, Lee, and Marino, 2001), a 1986 examination of the early career achievements of NIDA pre-doctoral trainees and fellows

(Clouet, 1986), and the National Institute of General Medical Sciences' 1998 review of its medical scientist trainees' career and professional activities. Using data from existing sources, numerous studies have addressed the characteristics of the research enterprise's workforce overall, e.g., the *Survey of Doctorate Recipients* and the *Graduate Students and Post-doctorates in Science and Engineering*, both conducted by the National Science Foundation. However, results for NIH trainees cannot be teased apart from these surveys' findings.

The NIH maintains databases of all applications and awards for NIH research and research training grants for every Institute. Hence, the NIH has the data available to measure NIH trainees' rates of NIH grant application and award. Unfortunately, these are essentially the *only* research training outcomes available in the NIH databases.

As the NIH points out, "Contributions to the scientific enterprise can occur through many different venues." In its most recent program announcement of T32 awards, the NIH stipulated that "evidence of a productive scientific career" could include

> a record of successful competition for research grants, receipt of special honors or awards, a record of publications, receipt of patents, promotion to scientific positions, and any other measure of success consistent with the nature and duration of the training received.

Hence, the range of possible outcomes indicative of a productive research career is wide, e.g., research grants from sources other than NIH, honors and awards, publications, patents, and scientific positions outside of NIH (such as faculty positions). Limiting an evaluation of NIH research training to only two outcomes—the extent to which NIH research trainees apply for and receive NIH grants—clearly represents much too narrow a measure of research productivity. Consequently, an adequate evaluation of NIH research training must reach beyond the data currently available in NIH databases.

In addition, appropriate research training outcomes to measure may differ depending on the trainees' stage of research training, as we expect different short-term outcomes from a pre-doctoral student with a T32 award (e.g., obtain doctoral degree and secure post-doctoral research position) than from an early-career researcher with a mentored K award (e.g., prepare an R01 application and secure an academic appointment). It was also expected that NIH research training

stakeholders might differ in what they consider to be meaningful outcomes of research training. Thus, a critical objective of the feasibility study was to determine the types of scientific contributions most valued by the Institutes' research training key stakeholders, particularly the three Institute Directors, and how to most feasibly measure these contributions.

NIMH, NIDA, AND NINDS RESEARCH TRAINING PROGRAMS

NIMH, NIDA, and NINDS have broad research portfolios supporting research "from molecules to managed care" and focus their research attention on related areas of neuroscience and behavior. To meet their scientific goals, each of the three Institutes supports a diversity of disciplines. For example, the NIMH supports psychiatry, psychology, social sciences, and neuroscience. Interdisciplinary training is therefore essential to ensure that future scientists are qualified to help these Institutes fulfill their goals. Not surprisingly, the types of departments receiving research training funds from the three Institutes overlap considerably. Investigators trained with support from one of these Institutes may well seek research funding from the other two Institutes. Thus, the missions of these three Institutes are linked and their research training goals are likewise interrelated.

NIMH, NINDS, and NIDA vary widely in their levels of investment in research training. NIMH devotes about 11.4% of its budget or $113.5 million to research training, which represents a very high level of research training investment (as a proportion of its overall budget) compared to most other Institutes. NINDS' investment in research training represents a smaller percentage (and smaller absolute dollar amount) of its budget, about 4.5% or $67.4 million, while NIDA devotes about 2.1% of its budget or $21.2 million to research training.[1]

The three Institutes also differ in how they have chosen to allocate their research training resources. NIMH invests nearly equally in NRSA and K awards, while NIDA and NINDS invest relatively more of their training resources in the NRSA program, though NIDA has also made a substantial investment in mentored K awards

There are several advantages to selecting NIMH, NIDA, and NINDS for the proposed evaluation. Taken together, these three Institutes present the opportunity to contrast the research training

[1] All budget estimates are based on fiscal year 2004.

outcomes for different levels of research training investment. Examining research training at three Institutes also increases the pool of trainees from which the evaluation can sample, allowing the evaluator to generate findings on subgroups of trainees, such as clinical research trainees. Finally, the three Institutes support a range of training mechanisms, offering the opportunity to compare outcomes of one training mechanism versus another within the same Institute and outcomes of a single training mechanism across Institutes.

2. Feasibility Study Research Questions and Methods

In this chapter we present the key research questions addressed in the feasibility study and the specific data collection methods and analysis approaches we employed to answer these questions

KEY RESEARCH QUESTIONS

We addressed three key research questions and their subparts in this feasibility study:

1. What are the **appropriate outcomes** of interest?
    - Identify and agree upon the desired results of the research training.
    - Are there recognized standards of performance agreed upon by all relevant stakeholders that could be used to assess research training success?
    - Of these, which are feasible to measure, and which will most efficiently reveal whether or not the Institutes' program goals are being achieved?
2. What **existing data sources** should be used to evaluate the program? What new data need to be collected?
3. If it is determined that there is a need for **primary data collection**, what is the best way to collect this data?

METHODOLOGY FOR ANSWERING THE KEY RESEARCH QUESTIONS

Due to the exploratory nature of the project, the feasibility study involved almost entirely qualitative methods, consisting primarily of targeted interviews with key stakeholders and individuals knowledgeable of the available and potential evaluation data sources. We also conducted a brief literature review of relevant research training evaluation studies (of which there were very few), and a small pilot study to test a primary data collection approach. *Exhibit 1* shows each of the feasibility study data sources and collection strategies mapped to the relevant research questions (questions 1 through 3 above).

Exhibit 1. Feasibility Study Data Sources and Collection Approaches

Research Question	Data Source	Data Collection Approach
What are the appropriate outcomes of interest?	NIMH Institute Director Dr. Thomas Insel NIDA Institute Director Dr. Nora Volkow NINDS Institute Director Dr. Story Landis	In-person interview
	Training Program Officers from NIMH, NIDA, and NINDS	3 In-person focus groups using protocol
	Institutional Training Program Directors/PIs receiving NIMH, NIDA, or NINDS training funds	15 Telephone interviews using interview guide
	Dr. Walter Schaffer, Acting Director, Office of Extramural Programs, Office of the Director	In-person interview
	Dr. Chuck Sherman, Acting Director, Office of Evaluation, Office of the Director	In-person interview
	Research training evaluations previously conducted	Literature review
	Literature on theory and practice of training scientists	Literature review
What existing data sources can be used?	Ms. Maria Bukowski, Office of Reports and Analysis, Office of the Director Dr. Bill McGarvey, Office of Extramural Programs, Office of the Director	In-person interview
	Dr. Chuck Sherman, Acting Director, Office of Evaluation, Office of the Director	In-person interview
	Dr. Michele Harmon, Senior Study Director, Westat (User of NIH grantee data)	In-person interview
	Non-NIH data sources including ▪ Association of Neuroscience Departments and Programs Survey ▪ Doctorate Records File ▪ Survey of Doctorate Recipients	Content review
What is the best way to collect the needed primary data?	Web and PubMed database	Web and PubMed search on purposive sample of known NIH trainees or grantees

What are the appropriate outcomes of interest?

To identify the appropriate outcomes of interest we analyzed all data sources by content analysis. We coded new themes as they emerged. For example, when multiple stakeholders independently raised the issue of the importance of protected time for post-doctoral research training, we identified this as an emerging theme and, in subsequent stakeholder interviews, asked follow-up questions and probed to explore this theme further. When an NIMH, NIDA, or NINDS Institute Director indicated a particular issue was important, it immediately became a key theme to explore with subsequent stakeholders. Thus, identifying the key outcomes to measure involved an iterative process by which we shared with one set of stakeholders the outcomes elicited from the previous set of stakeholders. With each successive set of stakeholders we further refined our concept of the key outcomes to measure.

What existing data sources can be used?

As a means of identifying existing data sources that could be used in the proposed evaluation, we conducted in-depth interviews with staff in the NIH Office of the Director who are knowledgeable of the types of data NIH collects on its research trainees and have extensive experience extracting these data from the NIH databases. We queried these individuals on the range of data elements available, how easily the data can be accessed, and data gaps and data quality issues. As part of our interviews with Principal Investigators (PIs) on NIH institutional training grants, we queried these PIs on the extent of their training programs' trainee outcome data efforts and their willingness to share these data with NIH in an evaluation of NIH research training. We also identified existing sources of comparison data outside of NIH and obtained reports and/or survey instruments on these data sources.

What is the best way to collect the needed primary data?

We developed a range of alternative data collection approaches and systematically evaluated each one for its feasibility in terms of data quality and completeness, efficiency of data collection, and burden on the individual or institution providing data. As part of our determining the most feasible way to go about collecting the needed primary data, we conducted a small pilot study involving Web and PubMed searches on a purposive sample of known former NIH trainees. Westat conducted a small pilot test of this electronic search approach with 30 individuals, including scientists in a range of behavioral health and biomedical fields relevant to the three Institutes' missions, as well as individuals outside of research but in science-related careers. Westat

researchers compared the data found from the Web and PubMed searches to the known career history and research outputs of each individual.

3. Feasibility Study Findings

In this chapter we present our feasibility study findings relevant to each of the key research questions. In separate memos and reports, we previously detailed the specific findings for each of the data sources in *Exhibit 1*. These memos and reports are shown in Appendices A through I. Here we synthesize those findings across data sources to answer the key research questions. In addition, we present feasibility study findings that while not directly related to the key research questions are nevertheless informative to the overall design of the research training evaluation. As we present the feasibility study findings we address their implications for conducting an evaluation of research training in the three NIH Institutes.

WHAT ARE THE APPROPRIATE OUTCOMES OF INTEREST?

Overview

While stakeholders expressed diverse views on the outcomes of interest in an NIMH, NIDA, and NINDS research training evaluation, all wanted to answer the basic question, *What do trainees do with their research training?* In particular, does the trainee go on to pursue a research career? If so, in what capacity? If not, does the trainee's career involve science at all? If the trainee achieves a research career, then nearly all stakeholders (including the NIMH, NIDA, and NINDS Institute Directors) would measure the individual's research productivity primarily by number and type of research grants and the number of publications in peer-reviewed journals. It is important to note that these main outcomes of interest all involve *long-term career outcomes*. That finding has at least two important consequences for the evaluation design:

- Measurement of long-term career outcomes necessitates a certain lag time after the conclusion of research training in order to allow sufficient time for the desired long-term outcomes to occur, e.g., faculty appointments, Research Program Grant (RPG) awards, and publications.
- If the three Institutes' research training mechanisms have the same long-term intended results and desired outcomes, then the same set of long-term outcome measures could be employed to evaluate research training across training mechanisms and Institutes.

Identify and agree upon the desired results of the research training

Every stakeholder readily generated a list of desirable (though not necessarily ideal) long-term outcomes for NIH research trainees that were the same across all research training award

mechanisms, but the contents of these lists varied widely across stakeholders. At one extreme, a minority of stakeholders viewed a career entirely outside of research, e.g., clinician, high school science teacher, or college administrator, as one of many desirable long-term results of NIH research training. At the other extreme, a small number of stakeholders stated that the only acceptable result of NIH research training is a career as an academic researcher.

The majority of stakeholders, though, including the three Institute Directors, generally saw as desirable (though not necessarily ideal) any career outcome in which the trainee was actively conducting research, whether as a PI or research team member; in NIH, academia, industry, or another setting; and regardless of funding source. It is worth noting that data to capture most of these desirable long-term career outcomes are not available in NIH databases or have not otherwise been collected.

Stakeholders, including the three Institute Directors, were in near complete agreement on the ideal or "gold standard" research training outcomes. The best possible research training outcome is an independent research career, either at NIH or in academia.

Interestingly, a substantial number of NIH research training program officers and other NIH staff, as well as a few of the NIH-funded institutional research training PIs, tended to view current NIH research training mechanisms as somewhat anachronistic in that the results they are intended to achieve, i.e., an academic researcher career, may be an unrealistic goal today, at least for some trainees. These stakeholders pointed out that obtaining a tenure-track academic appointment is so competitive today, with the number of qualified applicants far exceeding the number of available faculty slots, that this career outcome is not a realistic goal for every research trainee—there simply are not enough faculty slots to go around. A number of studies support this view (National Academy of Sciences, 2000; National Research Council, 2005).

If this scenario of limited academic research opportunities is accurate, then these circumstances should be factored into the research training evaluation. For example, if we assume that a substantial proportion of NIH research trainees will not secure academic research careers, are there other research training career outcomes that stakeholders consider desirable and which we would presumably want to measure in a research training evaluation? The answer appears to be "yes" on both counts. Stakeholders, including NIMH, NIDA, and NINDS Institute Directors, indicated a number of long-term career outcomes they considered desirable, including the conduct

of research almost anywhere. Capturing these desirable (though not ideal) outcomes would require primary data collection.

Finally, NIH research support was viewed by most as superior to all other sources, with some stakeholders outright disapproving of certain funding sources, such as foundations and industry. However, most stakeholders viewed most non-NIH funding sources favorably.

Determine if there are recognized standards of performance agreed upon by all relevant stakeholders that could be used to assess research training success

The vast majority of stakeholders, including the NIMH, NIDA, and NINDS Institute Directors, agreed that research training success is gauged over the long-term by research productivity. Likewise, there was widespread agreement (including among the three Institute Directors) on the indicators of research productivity, with the primary indicators of research productivity as follows:

- grant awards, especially NIH grants and, among these, RPGs in particular;
- academic appointments and tenure status;
- publications, particularly in peer-reviewed journals;
- citations;
- independent conduct of research in any setting;
- patents;
- scientific achievement awards; and
- leadership in scientific societies.

Stakeholders generally agreed that a research career marked by the above signs of research productivity clearly indicates a successful research training outcome. Of these indicators, the first three on the list—grant awards, academic appointments and tenure status, and publications (if in peer-reviewed journals)—are considered the most important of all and signal an ideal research training outcome, as the three Institute Directors and nearly all stakeholders agreed. Likewise, obtaining an RPG award in itself indicates an ideal research training outcome. Aside from NIH grants and research conducted within NIH, however, data on these indicators of research productivity are unavailable except through primary data collection.

The three Institute Directors, and most other stakeholders interviewed, also wanted to know about the career outcomes for trainees who do *not* attain these standards of research training success. If these individuals are not actively conducting research (as would appear to be the case if none of

the above indicators of research productivity are achieved), then what are they doing? Are they involved in science at all? Most stakeholders agreed that, while perhaps not viewed as a "research training success," a career involving science in some way, e.g., scientific editor, practitioner, or science teacher, was a more positive research training outcome than a position with no science involvement at all. A ranking of research training success could thus be constructed with gradations so as to distinguish among degrees of successful research training outcomes, e.g., an individual conducting research as part of a team in industry and who has published a few articles would be lower on the continuum of research training success than a PI with multiple RPG awards who is a tenured professor and has published on her research extensively. Again, primary data collection would be needed to measure and rank trainees' degree of research training success in this way.

Which of these performance standards are feasible to measure? Which will most efficiently reveal whether or not the Institutes' program goals are being achieved?

We have concluded that it is feasible—in terms of data quality and completeness, cost, time required to complete the evaluation, and burden on respondents—to measure all of the primary measures of research productivity listed previously, as well as classify the trainees' career outcomes (whether in a research, science-related, or other field), by career role, career setting, and source of research funding (if trainee is conducting research). For nearly all former NIH research trainees, these outcomes can be obtained from a combination of NIH database queries, Web searches, and PubMed searches. These measures, taken together, will most efficiently reveal whether or not the Institutes' program goals, in relation to research training, are being achieved. They will also answer most (although not all, as will be discussed later) of the questions the three Institute Directors would like to answer in an evaluation of their research training programs. Below we discuss how we came to this conclusion, feasibility issues around selecting the training mechanisms to evaluate, required time to complete the evaluation, measurement strategies, and cost estimates that support our feasibility assessment

Training Mechanisms to Include in the Evaluation

There was a clear consensus among the NIMH, NIDA, and NINDS Institute Directors and their research training program officers regarding the training mechanisms to include in the evaluation across the three Institutes: F31 (individual pre-doctoral), F32 (individual post-doctoral), T32 (institutional pre- and post-doctoral), and mentored K awards, including K clinical scientist awards. These awards comprise the bulk of the research training investments for the three Institutes.

Importantly, for each of these awards, the long-term desired result is the same: the development of independent researchers who engage in productive scientific careers.

As we noted earlier in this chapter, if the long-term desired results are the same across training mechanisms, then we can employ the same set of long-term outcome measures to evaluate all the research training mechanisms across the three Institutes. As we have concluded that this is indeed the case, different sets of outcome measures need not be developed for each of the training mechanisms, saving NIH substantial resources and time. Instead, the primary measures of research productivity (presented earlier in this chapter) and a classification of trainees' career outcomes can be used to measure trainees' long-term research training outcomes, regardless of the NIH training mechanism.

Each of the three Institute Directors was particularly interested in evaluating research training outcomes for NIH clinical research trainees with mentored K awards. After exploring the issue with NIH research training program officers, we have determined that the same set of long-term outcome measures can be used for clinical research trainees, although a lower overall volume of research productivity outputs is expected for this set of trainees. We anticipate this to be the case due to the nature of clinical research, e.g., research involving clinical populations typically takes much longer to conduct and is more expensive to conduct than basic research. Consequently, publications and grants would be expected to come to clinical researchers at a slower rate than for non-clinical researchers. In sum, it remains feasible to evaluate clinical research trainees using the same set of outcome measures but they should be analyzed separately.

Trainee Cohort Selection and Time to Complete the Evaluation

As we mentioned earlier in this chapter, measuring long-term career outcomes necessitates a certain lag time after the conclusion of research training in order to allow sufficient time for the desired long-term career outcomes to occur, e.g., faculty appointments, RPG awards, and publications. What is the optimal lag time to employ in the proposed evaluation?

According to a recently released report from the National Research Council (2005), the median age at which PhD researchers receive their first research grant is 42 years. This means that if the typical trainee is about 30 years of age when her NIH post-doctoral research training begins, we would not expect her to secure her first independent research grant for 12 years. Thus, the key research trainee career outcomes may not occur until a decade or more after the initiation of post-doctoral training, and even longer after the initiation of pre-doctoral research training. Clearly, a

longitudinal evaluation design in which current trainees are tracked for a decade or longer would not be feasible, as it would neither produce timely evaluation results nor be cost-effective. A more feasible evaluation design involves cohorts of trainees in the same training mechanism who began their training award at the same time, sufficient years ago to allow for the desired long-term outcomes to occur. In *chapter* 4 we detail this approach, including a sampling strategy that takes the required lag time into account.

WHAT EXISTING DATA SOURCES SHOULD BE USED TO EVALUATE THE PROGRAM? WHAT NEW DATA NEED TO BE COLLECTED?

We have identified two potential sources of existing data for the evaluation, NIH databases and outcome data collected by NIH-supported institutional training programs that are submitted to the sponsoring Institute. We have also identified existing sources of comparison data. Below we discuss the feasibility of employing these sources in the proposed NIH research training evaluation, followed by a brief discussion of the primary data that needs to be collected for an evaluation.

NIH Databases

The NIH Office of Reports and Analysis, within the Office of the Director, maintains (through its contractors) multiple databases of all NIH research and research training grant applications and awards for every Institute, as well as NIH appointments. There are three related databases:

- NIH Consolidated Grant Application and Fellow File (CGAFF)
- NIH Trainee and Fellow File (TFF)
- IMPAC I and II

Most importantly, the NIH databases can feasibly provide the evaluation's population of research trainees from which the evaluator would sample. In addition, data on NIH grants awarded to the research trainee, e.g., RPG awards, are available to the evaluation through these databases. However, as mentioned in *chapter 1*, these are the only research training outcomes available in the NIH databases. Moreover, these databases cannot reliably provide data on other types of involvement in NIH-funded research, e.g., Co-PI or Co-Investigator on NIH grants.

These databases are further limited in their usefulness to the evaluation in a number of ways. First, trainee contact information in the databases is very incomplete, e.g., incomplete mailing

addresses, missing telephone numbers, etc. For trainees receiving institutional training awards, the trainee contact information is not in the databases at all (only the PI's contact information is available). Furthermore, the contact information is not updated and thus will likely be out-of-date, especially for individuals who began their NIH research training ten or more years ago, as students, a particularly transient population. NIH staff familiar with these databases also expressed concerns about the amount of missing data in the databases.

Another major shortcoming of the databases, especially for trainees with common names, is that records of individuals with multiple NIH grants or appointments are not necessarily linked, which necessitates repeated queries in multiple databases and on multiple search terms. Likewise, individuals in the database do not necessarily have a unique identifier to link all their records, e.g., subsequent grants.

The NIH databases do contain contact information for trainees' mentors (for individual F and mentored K awards) and for trainees' training directors (the PI for institutional T32 awards), which could be useful to the evaluation if NIH wishes to contact the PIs or mentors for data collection. This represents a feasible use of the NIH databases, as discussed later in this chapter.

Outcome Data Collected by Institutional Training Programs and Submitted to NIH

All NIH-funded research training programs are required to collect data on their NIH-funded research trainees annually for up to ten years (from time of NIH trainee award) and submit the data to their sponsoring Institute. These data include the trainee's current position, institutional affiliation, source of support, and publications. All of the institutional PIs interviewed for this feasibility study were very willing to share the data with NIH as part of a research training evaluation. However, the three Institutes apparently do not maintain the data in any systematic way. As such, it is not available to the evaluation.

Comparison Data

A number of neuroscience training program directors suggested using the findings of the Association of Neuroscience Departments and Programs (ANDP) Survey as national comparison data against which to compare NIH pre- and post-doctoral research trainees. This survey boasts a response rate approaching 100% and provides data on trainees' (both NIH and non-NIH) career placement and funding sources. These comparisons would only be appropriate for research trainees in neuroscience programs.

Need for Primary Data Collection

There is a clear need for primary data collection in the proposed evaluation. The only outcomes available through existing data sources (the NIH databases) are receipt of NIH grants and NIH appointments. The vast majority of outcomes of interest are not captured by the NIH databases, such as research sponsored by other funding sources, academic appointments, research in industry and other settings, publications, and science-related career outcomes outside of research (e.g., science writer, practitioner, or policy maker). We conclude that data on these outcomes can only be obtained though primary data collection. The next section of this report explores the most feasible approach to collecting these primary data.

WHAT IS THE MOST FEASIBLE DATA COLLECTION STRATEGY?

Overview of Data Collection Alternatives

Various data collection approaches could be used to measure the long-term career outcomes of NIH trainees in each of the selected research training mechanisms. These data collection strategies include the following:

- survey trainees
    - by phone
    - by mail
    - by email or Web
- interview or conduct focus groups with trainees
- electronically search Web and archival data sources
- interview or survey trainee's research training PI or mentor

We briefly address the feasibility of each of these data collection approaches in terms of quality and completeness of the data, efficiency of data collection, and burden on the data provider.

Survey of Trainees

A survey of former NIH research trainees by phone, mail or email necessitates first obtaining contact information for the trainee in the desired mode (i.e., phone number for a phone survey, postal address for a mail survey, or email address for mail or Web survey). We found that trainee contact information in the NIH databases is very incomplete. For trainees with institutional training awards, the trainee contact information is not in the databases at all (only the PI's contact information is available). This is especially problematic given the large proportion of trainees with

institutional training awards in the three Institutes' research training portfolios. Furthermore, the contact information in the databases, when it is available, will be from five to nearly 20 years out-of-date and almost certainly no longer accurate.

Hence, current contact information would first have to be obtained for all trainees sampled for the evaluation. We have estimated that the evaluation will require a sample size of approximately 1,600 trainees (see *chapter 4* for discussion of trainee sampling estimates). Contact information for 1,600 trainees (called "tracing") can ordinarily be obtained through a data vendor at a reasonable cost (between $3,000 and $4,000). However, unless social security numbers are provided for each trainee, tracing would very likely produce no contact information for a substantial portion of trainees, probably 25% or more. The older the trainee's contact information is, the lower the likelihood of obtaining a current address or telephone number. "Misses" are particularly likely for pre-doctoral trainees, whose addresses will be from the late 1980s and early 1990s (to account for the required lag time). It is Westat's assessment that unless trainees' social security numbers are provided to the evaluator, usable contact information cannot be obtained for a sufficient number of NIH trainees to make any approach requiring trainee contact feasible, including survey or interview via phone, mail, email, or Web.

Interview or Conduct Focus Groups with Trainees

In addition to the cost of traveling and meeting trainees, which would run well into the millions of dollars for the entire sample, this approach is not feasible because the evaluator would first have to obtain trainees' contact information, as described above, which we have already determined is not feasible. If a small sub-sample of trainees (approximately 50) were selected for telephone interviews in order to provide in-depth responses to special topic research questions, this could be a feasible approach, but not for a larger sample and not to answer the key research questions on long-term career outcomes.

For example, if NIH wanted to learn more about the barriers to a successful research career, the evaluator might interview over the phone 50 trainees who did not pursue a research career. While contact information would need to be obtained for the 50 trainees, the task would be feasible for this small number of trainees if it involved, for example, only individual post-doctoral awards (individual trainee contact information would be in the NIH database, but institutional trainee award contact information would not; post-doctoral awardees would have more recent contact information than pre-doctoral awardees). As an added component to the evaluation, this would take

approximately 400 hours to carry out and cost approximately $30,000, depending on the nature and length of the interview. This data collection task would likely require OMB clearance, but if the OMB clearance package were prepared and submitted early in the evaluation, this evaluation component would remain feasible.

Electronically Search Web and Archival Data Sources

We believe that the most feasible approach to measuring the outcomes of interest is a combination of searches in the NIH databases, PubMed database, and on the Web.

- The **NIH database contains data** on NIH grant applications and awards as well as NIH appointments
- The **PubMed database** has a comprehensive catalogue of publications on biomedical and behavioral research. In addition, the evaluator can use PubMed's MESH (Medical Subjects Headings) thesaurus to categorize the publication content by research area. For example, if NIH would like to know if clinical research trainees are now doing clinical research, MESH terms indicating clinical research could be selected to categorize publications as such.
- Generic **Web searches** can provide evidence of all the indicators of research productivity, including academic and other professional appointments, grants, leadership roles in professional science, etc.

In *chapter 2* we described our conduct of a small pilot study involving Web and PubMed searches. While not a rigorous evaluation of this data collection method, the results were nevertheless very promising. In all but two cases (one a new researcher and the other not engaged in a science career at all), fairly complete and accurate "dossiers" on each individual could be constructed, providing a clear picture of the individual's research and scientific involvement, especially for scientifically productive individuals. That is, number and types of publications, academic affiliations, and type of career and career setting were obtainable for all but the two cases noted.

We conclude that the "false positive rate" with this approach, i.e., wrongly crediting a trainee with indicators of research productivity, is very low. The "false negative rate," i.e., failing to credit a trainee with indicators of research productivity, appears somewhat higher but still fairly low. In other words, if a trainee actually has indicators of research productivity, these indicators are very hard to overlook in Web and PubMed searches. False negatives appear most likely to occur when the

trainee has changed his or her name (most often upon marriage) subsequent to beginning research training or when the name is a common one, e.g., “Robert Smith.”

We estimate that this data collection task alone would take about 4,800 hours to complete, including data entry, at a cost of approximately $350,000. An evaluation using this strategy would take about 18 months to complete.

The major shortcoming of this approach is that trainees with common names and individuals who have changed their name since their NIH research training may not be found through these methods. (The NIH databases do not link the records of an individual with a name change.) The evaluator will readily recognize a common name, e.g., John Smith. If a trainee changes his or her name, though, that will not be obvious to the evaluator. However, when the PubMed and electronic search turn up very little, there may be a name change involved. Or the individual may have made little impression on the scientific world. It is hard to know which situation is the case without additional information. An option that would help to fill in those data gaps would be to collect data from the trainee’s research training PI or mentor via telephone interview or survey. We describe this supplemental data collection approach in the next section.

Contact Trainee’s Research Training PI or Mentor

Used as a supplemental data collection strategy only when the above combination of NIH databases, Web search and PubMed search gives no or very little indication of the trainee’s career outcome, contacting the trainees’ PI or mentor is a reasonable way to obtain a limited amount of outcome data on trainees. As the first attempt to collect data on trainees’ long-term career outcomes, this strategy is not efficient in terms of burden on the PI or mentor, completeness of the data he or she can provide, and cost. However, as a back-up or *supplemental* data source, when the first-line data collection strategy (i.e., combination of NIH databases, Web, and PubMed) falls short, contacting the PI or mentor may be a feasible approach

Research trainees’ NIH training program PIs and mentors generally know quite a bit about their former trainees’ career outcomes (as we learned from interviewing them), particularly the trainees’ faculty appointments and grants, leadership in scientific societies, and prestigious awards. The PI or mentor is likely to know trainees’ new names after marriage, especially if the trainee is actively involved in research. Likewise, given a trainee with a common name, the PI or mentor may be able

to provide enough information on the individual for the evaluator to discern which "John Smith" is the sampled NIH research trainee.

The NIH databases contain contact information for the trainee's mentor (for individual F and mentored K awards) and for the trainee's training director (the PI for institutional T awards). However, the databases will almost certainly have more recent contact information for the institutional PI or mentor than for the trainees themselves because (1) the PIs/mentors were already established faculty members (rather than transient students) at the time of the trainee's award and many will have remained on the faculty, and (2) many of the PIs/mentors received subsequent NIH grants and research training grants, served as mentors for other F or K awardees, or served on an NIH Institute Council. Even if the NIH databases do not provide current contact information for the PI/mentor, as established academic researchers, they will be easily locatable either through a data vendor or via Web search. (It can be expected that a minority proportion of PIs/mentors will have retired or died, particularly those of sampled pre-doctoral trainees.) On balance, we believe obtaining contact information for approximately 10% or 160 research training PI/mentors is feasible and of relatively modest cost.

The task of contacting PIs/mentors and obtaining limited outcome data for 160 trainees would take approximately 500 hours, excluding data entry, and cost approximately $36,000. A written request sent to the PI/mentor, rather than a telephone contact, would reduce the effort and cost required to complete this task only marginally. However, a written request would limit the number and types of questions that could be asked of the PI/mentor with no opportunity for clarification or immediate follow-up. Thus, we believe that a telephone contact, even if very brief, would be a more efficient strategy for obtaining limited outcome data from PIs/mentors than would a written request.

There are two significant drawbacks to obtaining more extensive outcome data from the PI/mentor using this approach. First, it is unlikely that the PI/mentor will be able to provide complete data on the indicators of research productivity, as the trainee may have been last seen by the PI/mentor from 5 to over 15 years ago. The PI/mentor cannot be expected to know or report on the full range of indicators of long-term research productivity. This strategy is best viewed as a means of providing "seed data" on an otherwise "missing" trainee; the evaluator would expect to use the seed data, e.g., new name, place of employment, type of career pursued, in subsequent searches of the Web, NIH databases, and PubMed.

Second, some PIs/mentors serve many research trainees. This is particularly the case for the PIs on NIH institutional training grants, who may have had six or more NIH research trainees on a single training grant each year over many years. It may become burdensome for such a PI/mentor to provide data on the indicators of research productivity for multiple NIH trainees. However, this concern is somewhat mitigated by the knowledge that for at least two of the three Institutes in the evaluation, sampling will be used to select trainees on pre- and post-doctoral institutional training awards because of the large numbers of trainees receiving these awards. The sampling reduces the likelihood that any one PI/mentor would be asked to report on a large number of trainees.

ADDITIONAL FINDINGS INFORMATIVE TO THE EVALUATION

In this section we present selected feasibility study findings that are informative to the evaluation design though not necessarily directly related to the key research questions addressed above.

Individual versus Institutional Awards

Many of the PIs for NIH institutional training grants made a point of comparing trainees on institutional training awards, most often the T32, with trainees on NIH training grants awarded directly to the individual, most often the F31 or F32. These PIs held the individual-level awards in much higher esteem than the T32 awards, typically observing that trainees "really have to compete" for the individual-level awards, while someone with a T32 "didn't have to do anything" to get it. While the PIs noted that they had high expectations for all trainees in their programs, they were especially impressed by trainees with individual NIH awards and believed such individuals had very much distinguished themselves as having the potential to be independent researchers. Consequently, the proposed evaluation addresses a research question comparing trainees with individual and institutional training awards at the same stage of training.

Selection Effects

The major methodological concern in evaluating the effect of NIH research training is how to separate the selection effects of receiving an NIH training grant from the impact of the training grant itself, i.e., if students selected for an NIH training grant are presumably more qualified at the start of training than those not selected, then how can differences in outcome be attributed to the impact of the NIH research training? Might the selected students have been just as successful *without* the NIH training grant? This selection effect is also at work at the training program level.

Those training programs selected to receive NIH research training grants are likely different at baseline from similar research training programs who either do not apply for NIH research training grants or apply but are rejected. These circumstances make finding an appropriate comparison group for NIH research trainees problematic. Consequently, the proposed evaluation design does not involve data collection from comparison groups (although comparison using existing national data sources is presented as an evaluation option).

Variation in Use of Institutional Research Training Grants

PIs and training programs vary considerably in how they use NIH institutional training grants (e.g., the pre- and post-doctoral T32). For example, some training programs single out NIH institutional research training grant recipients for special attention and enrichment opportunities, whereas others treat all trainees in the program the same regardless of funding source. In the latter programs in particular, all training program participants reap the benefits of the NIH institutional training grant, whether they are the actual grant recipients or not. For example, a number of PIs of T32 training grants indicated that all students in their program—whether sponsored by a T32, faculty R01, or some other funding—receive essentially the same training experience, with only minor variations due to funding source.

The most substantial benefits of NIH institutional training grants may go to the larger training program and are enjoyed by all trainees regardless of funding source. For example, T32 training programs are required to have an ethics course, which in most programs is offered to all the trainees, not just those on the T32. Thus, in evaluating trainees on NIH institutional research training grants, we expect that differences among training programs contributes substantial variance to the individual-level outcomes, while differences between T32 trainees and non-T32 trainees are minimal in many programs. Again, these circumstances make selection of an appropriate comparison group difficult and contribute to our decision not to collect comparison group data.

Research Trainee Career Outcomes Reported by Institutional PIs

The PIs of NIH institutional training programs (primarily T32 awards) that we interviewed for this study reported that between 75% and 90% of their post-doctoral trainees eventually make careers in academic research. Those percentages varied depending largely on the type of training program, with programs in neuroscience fields more likely to place its trainees in faculty research positions. These outcomes are much more positive than those reported for the biomedical field as

a whole (National Research Council, 2005). We speculate that the higher rate of faculty appointments for post-doctorates in these feasibility study programs is due to the highly selective nature of these programs. The feasibility study training programs are distinguished first by their receipt of an NIH research training grant. Second, NIH selected these programs' PIs for the feasibility study due to the PIs' extensive involvement in NIH research training. Thus, there is reason to expect that these are exceptionally well-run training programs whose trainees may be afforded research training benefits unavailable to trainees in other programs. Yet, it is also possible that, to some degree, the programs may have over-stated their programs' success and/or ignored or forgotten about their programs' "failures." Consequently, we expect that substantially fewer than 75% of the trainees in the proposed evaluation will be in academic positions.

Research Training Strategies

Some stakeholders, including two of the three Institute Directors, would like the evaluation to produce data on the optimal research training strategies, as well as the outcomes indicative of research productivity discussed earlier. One Institute Director posed very specific questions regarding the administration of research training programs, e.g., What is the optimal administration of research training strategies? Should training for Fs and Ks be organized within the Institute, jointly in cross-Institute training programs, or as currently administered? The same Institute Director would like to answer questions such as, What is the optimal combination of research training award mechanisms? What is the optimal amount for a training grant? What is the right amount of protected time for various K awards?

Unfortunately, these questions are beyond the scope of the proposed research training evaluation. However, the proposed evaluation, which primarily addresses the outcomes of research training, would be essential to conduct before these more nuanced questions on the best research strategies could be addressed. The proposed evaluation will provide a wealth of data on the outcomes for each training mechanism and establish a baseline of performance for each mechanism. Institute Directors will then have the data they need to assess the extent to which these training mechanisms are producing the desired results for their Institutes. Outcomes among various training mechanisms and Institutes can then be compared for relative efficiency.

Other stakeholders, including a different Institute Director, were especially interested in measuring various aspects of mentorship and the role it plays in developing a successful research career. However, mentorship is a construct that would require a great deal of further conceptualization before we could begin to measure it. In addition, relating mentorship to specific research training

outcomes would be an enormous undertaking, constituting a study in itself. As such, this is also beyond the scope of the proposed evaluation.

There was also interest in examining to what extent career paths and outcomes vary by gender and identifying the barriers to an independent research career. It will be feasible in the proposed design to examine outcomes for trainees by gender, at least for certain training mechanisms with large enough sample sizes. If the Institutes would like to examine the barriers to an independent research career, a fairly small sample of trainees (approximately 50) could be sampled and direct contact with trainees initiated. This would be particularly feasible to carry out if the training were fairly recent, so that their contact information was more likely to be current and memories of the barriers fresh.

External Factors

Certain factors external to the trainee and the research training process may influence the outcome of research training, particularly workforce trends and the demand for researchers in particular areas. For example, the apparent trend of more faculty applicants than available faculty positions reduces the likelihood of securing an academic research position. At the same time, researchers in specific training areas may be in demand, as currently seems the case for the neuroscience research trainees. Time trends such as these could be partially accounted for in the proposed evaluation by comparing cohorts of NIH research trainees at the same stage of training who being training during different time periods. However, this design is probably not feasible given the substantial additional data collection costs incurred in adding cohorts. For example, we estimated that the data collection task alone would take approximately 4,800 hours and cost $350,000 for one cohort of trainees. Adding a second cohort would essentially double the hours required and cost of data collection.

4. Proposed Evaluation Design

In this chapter we outline the proposed evaluation research questions, present a conceptual framework of NIH research training, detail key outcomes to be measured, and describe the proposed data collection and analysis approaches.

RESEARCH QUESTIONS

Based on the information needs of the NINDS, NIMH, and NIDA Institute Directors, we propose the research questions listed below to drive the evaluation of the Institutes' research training programs. These questions are broadly grouped into the following topic areas: career outcomes, research productivity outcomes, research training progress, individual versus institutional awards, and comparisons among Institutes. Unless otherwise noted, "trainees" refers to NIH research trainees in NIMH, NIDA, or NINDS on T32 (pre- and post-doctoral), F31, F32, or mentored K (K01, K08, K12, K22, K23, and K25) awards.

Career Outcomes

What are NIH research trainees' career outcomes, given sufficient time to establish a research career?

- For each training award type, what proportion of trainees is conducting research independently? In what settings are they conducting research?
- For each training award type, what proportion of trainees has each of the following career roles: research, administration, science-related role, or a non-science-related role?
- For each training award type, what proportion of trainees goes into each of the following settings after completing their training: NIH, other government, industry, academia, non-academic research institute, health care, other?
- For each training award type, what proportion of trainees has an academic appointment? What proportion of those appointments is tenured? What is the Carnegie Classification[2] of the institution?

[2] The Carnegie Classification is a typology of all degree-granting colleges and universities in the United States. Every institution has a Carnegie Classification. The Classification of the institution at which a trainee holds a faculty appointment can be viewed as an indication of the institution's commitment to research.

Research Productivity Outcomes

- For each training award type, of the trainees in research roles, what proportion of trainees receives research funding from the following sources: NIH, other government sources, private foundations, industry, other?
- For each training award type, what proportion of trainees obtains at least one Research Project Grant (RPG)?
- For each training award type, what proportion of trainees publishes in academic journals? For trainees publishing, what proportion publishes in peer-reviewed journals? What proportion publishes in the content area in which they were trained or one closely related?
- For each training award type, what proportion of trainees conducts independent research outside of NIH funding sources?
- How do these outcomes differ for trainees with clinical research training awards?

Research Training Progress

How do NIH research trainees progress through NIH research training?

- What are the typical trainee profiles for NIH research training mechanisms? For example, does a trainee typically move from the T32 to the mentored K to an R01, or does the trainee go directly from the T32 to the R01?
- Does this profile differ for clinical research trainees?
- How do the trainees differ across the three Institutes in their use of NIH research training mechanisms?
- To what extent do research trainees use multiple NIH post-doctoral training awards, e.g.,T32s?
- Does age of research trainee at time of training award affect the trainee profile?
- Do other variables, such as research trainee gender, affect the trainee profile?

Individual versus Institutional Awards

Do NIH research training awards that require individual-level competition produce better outcomes than those involving only institutional-level competition?

- How do T32 pre-doctorals and F31s who start their training awards at the same time compare in their career outcomes.
- How do T32 post-doctorals and F32s compare?

Comparisons among Institutes

- For all of the above, how do trainees in NIMH, NINDS, and NIDA differ?

CONCEPTUAL FRAMEWORK OF NIH RESEARCH TRAINING

The conceptual framework of research training at NIMH, NIDA, and NINDS that will drive the evaluation is presented in *Exhibit 2*. The conceptual framework depicts the inputs into the research training process and external factors affecting this process, the intended training activities, the expected training process goals, and the intended intermediate and long-term goals of the research training. As the exhibit shows, the intended long-term goals of NIH research training are for the trainee to:

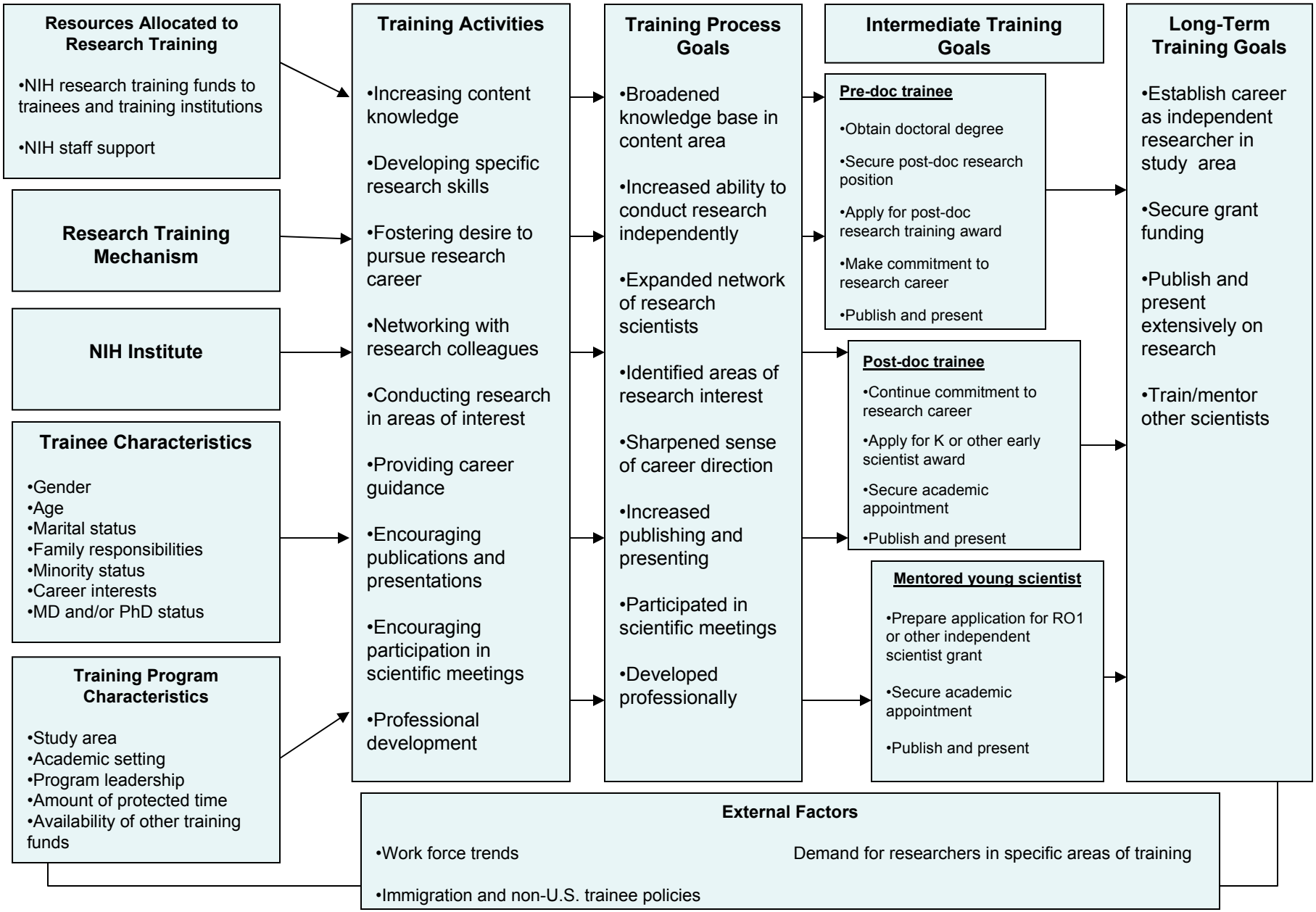
- Establish a career as an independent researcher in the indicated study area,
- Secure grant funding,
- Publish and present extensively in the research area, and
- Train and mentor other scientists to enter research.

As can be seen in the conceptual framework, while intermediate training goals vary depending on the trainee's stage of training (i.e., pre-doctoral, post-doctoral, or mentored young scientist), the training activities, training process goals, and long-term training goals are conceptually the same. The proposed evaluation focuses almost exclusively on the long-term outcomes of training, which are the same for all trainees, regardless of training mechanism or Institute, as discussed in *chapter 3*.

TARGET POPULATION

Research trainees awarded T32 (pre- and post-doctoral), F31, F32, or mentored K awards (K01, K08, K12, K22, K23, and K25) by NIMH, NIDA, and NINDS constitute the target population for this evaluation and the unit of analysis is the individual trainee.

Exhibit 2. Conceptual Framework of Research Training at NIMH, NIDA and NINDS



KEY VARIABLES

The key dependent variables in this evaluation of NIH research training are the performance measures relating to the long-term goals of training, i.e., outputs indicative of research productivity and other science-related activity and trainees' career outcomes. *Exhibit 3* shows a typology of career outcomes to be used in the evaluation to classify trainees' career outcomes along three dimensions: career role, career setting, and research funding (if conducting research). This typology reflects the range of likely career outcomes for research trainees, both within and outside of the NIH (including industry), which stakeholders would like to measure. *Exhibit 4* lists another set of key variables to be measured in the evaluation, specific outputs indicative of research productivity and science-related activity. The key independent variable in the proposed evaluation is the type of research training mechanism, but depending on the research question, other independent variables are the Institute (i.e., NIMH, NIDA, or NINDS), type of research training (i.e., clinical versus non-clinical and individual versus institutional), and trainee gender.

Exhibit 3. Dimensions of Career Outcomes for NIMH, NIDA, and NINDS Research Trainees

Dimension 1: Role

- Research
  - Independent, e.g., Lab Chief, PI, Co-PI, or Scientific Director
  - Supporting research role, e.g., Investigator, member of a research team
- Administration (does not involve conduct of research)
- Science-related Role (does not involve conduct or administration of research)
  - Practitioner
  - Clinical Trainer
  - Scientific Editor
  - Faculty Member
  - Science Teacher
  - Policy Maker
  - Other
- Non-science-related role (involves science only marginally or not all)

Dimension 2: Setting

- NIH
- Other government
- Industry
- Academia
- Non-academic research institute
- Health Care
- Other

Dimension 3: Source of Research Funding (if primary role is Research Conduct)

- NIH
- Other Government Source
- Private Foundation
- Industry
- Other

Exhibit 4. Outputs Indicative of Research Productivity and Science-related Activity

- Grants and research funding awards (all types)
    - Categorize by funding source
- NIH grant applications
- Academic appointments
    - Carnegie Classification of University where appointed
    - Tenure status
- Scientific journal publications
    - Categorize by peer-review status of journal
    - Categorize by content
    - Number of publications
- Citation history
- Other science-related publications, includes
    - book chapters
    - articles in popular press
    - training materials
    - reports
- Patents
- Leadership positions in scientific societies
- Professional presentations
- Scientific achievement awards

RECOMMENDED DATA COLLECTION APPROACH

In this section we summarize our recommended data collection approach (having already reviewed its feasibility for the proposed evaluation in *chapter 3*), which is based on considerations of data quality and completeness, efficiency of the data collection approach, data collection cost, and burden to the data provider. The recommended approach is represented by the *core study*, described below, which in itself will adequately answer the evaluation research questions. In addition to the core study, we present three options that the Institutes may or may not want to add the core study: (1) an optional supplemental data collection approach to enhance the quality and completeness of data collected in the core study, (2) an optional evaluation component to address a separate set of research questions (e.g., barriers and enhancers to a successful research career), and (3) an optional evaluation component incorporating comparison groups. We summarize each of these approaches below.

Core Study

The core study involves the following three data sources:

- NIH databases
- PubMed database
- Web search

Using these three data sources, which we described in detail in *chapter 3*, all the research questions posed earlier in this chapter can be feasibly addressed. Data on the research trainees to be sampled and on their outcomes of NIH grants and appointments are contained in the NIH databases. Searching PubMed will provide data on trainees' scientific publications, the content of these publications, and, in many cases, trainees' institutional affiliations (within the publication text). The Web search will reveal data on career role and setting outcomes as well as the outputs indicative of research productivity and science-related activity shown in *exhibit 4*.

A key feature of this core study is that comprehensiveness of data, in the absolute sense, is not necessary to answer the evaluation questions. That is, if some of the above mentioned data elements are missed, this will not necessarily influence the findings. For example, if the evaluator collects data indicating the trainee published 20 peer-reviewed journal articles in her field of research training, it is inconsequential to the evaluation if the evaluator "fails" to collect data on the other dozen articles the trainee has also published. For the purposes of the evaluation, the trainee's high scientific publication rate has already been firmly established. Likewise, if the

evaluator documents that the trainee held a tenured position in a top-rated research university (by Carnegie Classification), but does not capture data on a subsequent academic appointment, this "missing" data does not materially influence the evaluation findings at all. The key outcome of tenured academic position has already been captured.

As we detailed earlier in *chapter 3* when we described our successful pilot study of this approach, this core study strategy carries a very low false positive rate, i.e., it is very unlikely that trainees will be wrongly credited with career outcomes or indicators of research productivity that do not belong to them. We expect the false negative rates, i.e., failing to credit trainees with outcomes that actually belong to them, to be somewhat higher but still fairly low. False negatives appear most likely when the trainee has changed his or her name (most often upon marriage) or when the name is a common one, e.g., "Robert Smith."

Optional Supplemental Data Collection Approach

This optional supplemental data collection approach, which is not part of the core study, could be added to the evaluation to enhance the quality and completeness of data collected in the core study and reduce the likelihood of false negatives, as discussed above. When a sampled trainee comes up "missing," i.e., no data is available regarding the trainee's career outcomes or indicators of research productivity, there are two possible explanations: (1) a "false negative," i.e., the trainee has engaged in research or science-related activities but they are not high-profile activities and thus are not accessible on the Web, or (2) the trainee has not engaged in such activities and the "missing data" actually represents a "true negative," i.e., the trainee truly has no indicators of research productivity and no indication of a research or science-related career. We must note here that in our pilot study we found that it is very difficult for an individual actively engaged in research to avoid detection using the core study approach (case 1). However, when an individual has changed his or her name subsequent to the start of research training, such a false negative is likely.

The evaluator who finds no outcomes for a trainee will not know whether this represents a false or true negative. To resolve this issue, an optional data collection approach could be used: contacting the trainee's former research training PI using the contact information available in the NIH databases. We described this strategy in detail in *chapter 3* and deemed it feasible if used only "as-needed," as in the circumstance just depicted. Based on our interviews with PIs and mentors, there is a good chance they will know about trainees' name changes and be willing to

provide the new name, know what trainees are doing in their career, and generally have a sense of whether the individual pursued a research career or not, and if so, where. The PI/mentor might also be contacted when a trainee has a common name and the evaluator cannot discern what outcomes are rightfully associated with the sampled trainee versus other individuals with the same name.

Optional Evaluation Component to Address Additional Research Questions

If the three Institutes so desire, data could be collected directly from a sub-sample of trainees in response to an additional research question that requires direct communication with trainees. For example, the Institutes have expressed an interest in learning more about barriers and facilitators to an independent research career, especially the role of mentorship. The evaluator could sample a small number of trainees (e.g., 50) who did not go on to research careers, as identified through the core study, and survey or interview them by phone on the selected topics. In *chapter 3* we discussed the infeasibility of obtaining accurate contact information for all trainees in the core study. However, it may be possible to achieve acceptable contact and response rates for 50 trainees, which would make this approach feasible.

Optional Evaluation Component Using Comparison Groups

As discussed at length in *chapter 3*, appropriate comparison groups for the proposed evaluation are not readily available. However, NIH research trainees could be compared to national samples, but only on a narrow set of outcomes. Most of these outcomes are short-term, rather than long-term career outcomes. For example, the Doctorate Records File, which is based on the Survey of Earned Doctorates, reports doctorates' employment or post-doctoral plans at time of graduation. The Association of Neuroscience Departments and Programs (ANDP) has offered to share their survey findings with the three Institutes for use in the proposed evaluation. The ANDP survey reports on the placement of new doctorates and post-doctorates and their sources of funding. These comparisons would only be relevant for trainees in neuroscience programs. If the three Institutes have a strong interest in making any possible comparisons between their research trainees and national samples, this optional component would be feasible. The resources required to carry out this component would be minimal (approximately 40 hours and costing about $2,500), although the findings would only indirectly relate to the evaluation research questions.

Summary of Core Study and Optional Data Sources and Collection Approaches

Exhibit 5 summarizes the data sources and collection approaches in both the core study and options just described.

NEW DATA COLLECTION INSTRUMENTS

The evaluation will require the development of some basic data collection instruments to carry out the core study of the proposed design. The following data collection instruments will be needed for the core study:

- **NIH database extraction sheet** to collect NIH research grant and grant application and award outcomes from the NIH databases as well as
- **Web search extraction sheet** to collect career dimensions (see *exhibit 3*) and indicators of research productivity (see *exhibit 4*) from Web searches
- **PubMed extraction sheet** to record publications

In addition, if the Institutes choose to include the options described in the previous section, the following instruments will also be needed:

- **PI Interview Guide or Survey** to collect data from the trainee's NIH research training PI/mentor
- **Trainee Interview Guide or Survey** to collect data directly from trainee on topics selected by the Institutes

CLEARANCE REQUIREMENTS

If a contractor is used to conduct the proposed evaluation, it may be necessary for the contractor to obtain special permission from the NIH Office of the Director to use the NIH databases for (a) drawing the sample, (b) obtaining any available identifying information on sampled trainees (if the NIH permits) in order to facilitate the Web search, (c) identifying any subsequent NIH grants, applications, or appointments, and (d) contacting the trainee's NIH research training PI/mentor if this supplemental option to the evaluation is selected. Likewise, if NIH chooses the optional evaluation component of contacting trainees directly to systematically survey or interview them, then clearance from the U.S. Office of Management and Budget (OMB) will be required. Finally,

Exhibit 5. Core Study and Optional Data Collection Approaches

	Data Source and Collection Mechanism	Outcomes Available	Comment
Core Study	Data extraction from NIH databases (TFF, CGF, and IMPAC)	▪ NIH grants and awards ▪ NIH PI status ▪ NIH applications	NIH non-PI research involvement partially available Cannot obtain adequate range of outcomes from this data source
	NIH databases, Web and PubMed searches Employ combination of Web site review, database extraction, and literature review (of publications obtained using these methods)	▪ NIH grants and awards ▪ Publications ▪ Citation history ▪ Patents ▪ Academic appointments ▪ Position and job setting	Cost-effective way to collect data on a number of relevant outcomes Pursue "misses" further by contacting the PI/mentor of institution Outcomes partially available: ▪ Non-NIH research funding ▪ Position /setting outside field ▪ Presentations ▪ Leadership in professional or scientific associations ▪ Scientific awards
Supplemental Option	Contact trainee's NIH training PI and/or institution in which received NIH-funded training May lead to contacting trainee.	▪ Career role and setting, whether or not pursued research career ▪ In most cases only a limited range of outcomes available. ▪ Highly variable, but potential for complete dossier	Very labor intensive as a primary data collection strategy; consider using as a supplemental strategy to Web search NIH training award PI/mentor contact data available in NIH database Good strategy to find trainees who changed name or trainees with common names
Trainee Contact Option	Administer survey or structured telephone interview to trainee. Must trace trainee and make personal contact.	▪ Positions /settings, current & past ▪ Non-NIH research activity ▪ Barriers to research career, role of mentorship ▪ All information to build a trainee dossier	Very labor-intensive as a primary data collection strategy. Feasible only as a supplement to Core Study
Comparison Option	For comparison: Doctorate Records File (DRF), based on Survey of Earned Doctorates (SED)/Survey of Doctorate Recipients (SDR) Employ database extraction and/or document review	▪ Doctorate degree type, date, and institution of new PhDs ▪ Employment or post-doctoral plans ▪ Time to degree ▪ Demographics and education history (process variables)	Difficult to obtain adequate comparison group matching trainees because of baseline differences. Can provide gauge of secular trends as a context interpreting evaluation findings, but only on a fairly narrow range of short-term outcomes.
	For comparison in neuroscience only: ANDP Survey of Neuroscience Grad, Post-doctoral, and Undergrad Programs Employ database extraction and/or document review	▪ Placement setting of new PhDs, post-doctorals ▪ Source of funding for PhD, post-doctorals	

the evaluator must comply with all Privacy Act Requirements and obtain IRB approval for the evaluation.

DATA INTEGRITY

Strategies should be developed and implemented to ensure that the evaluation data collection is accurate and complete. All data collection instruments and procedures should be pilot tested prior to their use in the field. In particular, inter-rater reliability should be established for the protocols used for the Web searches, PubMed searches, and NIH searches. Likewise, protocols should be established for each of the six data collection sources and approaches shown in *exhibit 5*. Data coding schemes for all data elements must also be developed. Finally, the evaluator should develop and implement procedures for training and monitoring the data collectors as well as for storing and use of the data collected.

SAMPLING

The proposed sampling plan involves separate cohorts of same-stage-of-training trainees for each research training mechanism from each of the three Institutes. In addition, each cohort is comprised of trainees who began training in the same 2-year period. That 2-year "training start" period has been chosen such that the lag time between the start of training and the evaluation allows trainees sufficient time to achieve the desired long-term outcomes of research training, as discussed in *chapter 3. Exhibit 6* illustrates the sampling process, which will be repeated for each of the three Institutes

Exhibit 6. Sampling by Stage-of-Training, Training Mechanism, and Year Began Training

Stage of Training	Training Mechanisms	Time Cohort (Year Began Training)
Pre-doctoral	T32, F31	1990, 1991
Post-doctoral	T32, F32	1993, 1994
Early Career	Mentored K (various)	1996, 1997

Using preliminary estimates of the number of trainees in each of the cohorts, we anticipate that sampling can be used for the T32 pre- and post-doctoral trainees because of the large numbers of trainees in these cohorts. By sampling trainees in a cohort, rather than selecting them all, data collection can be significantly reduced and the evaluation can be completed more quickly and at lower cost.

We have constructed the preliminary sample to provide estimates on the outcomes for each training mechanism cohort with a 95% confidence interval and 5% or lower sampling error. The size of this preliminary sample, including all cohorts across the three Institutes, is approximately 1,600 trainees.

DATA ANALYSIS

For the majority of research questions, descriptive statistics will be used. For example, the evaluator will use descriptive statistics to report, for each award type, the proportion of trainees engaged in science-related fields that do not involve conducting research. For all research questions involving comparisons, statistical hypothesis testing will be used. *Exhibit 7* shows the analysis method that will be used to answer each research question.

Exhibit 7. Methods of Analysis for Research Questions

Research Questions	Analysis Methods
Career Outcomes What are NIH research trainees' career outcomes, given sufficient time to establish a research career?	
▪ For each training award type, what proportion of trainees is conducting research independently? In what settings are they conducting research? ▪ For each training award type, what proportion of trainees has each of the following career roles: research, administration, science-related role, or a non-science-related role? ▪ For each training award type, what proportion of trainees goes into each of the following settings after completing their training: NIH, other government, industry, academia, non-academic research institute, health care, other? ▪ For each training award type, what proportion of trainees has an academic appointment? What proportion of those appointments is tenured? What is the Carnegie Classification of the institution?	descriptive statistics

Exhibit 7 continued

Research Questions	Analysis Methods
Research Productivity Outcomes	
▪ For each training award type, of the trainees in research roles, what proportion of trainees receives research funding from the following sources: NIH, other government sources, private foundations, industry, other?	descriptive statistics
▪ For each training award type, what proportion of trainees obtains an RPG award?	descriptive statistics
▪ For each training award type, what proportion of trainees is publishing in academic journals? For those trainees publishing, what proportion publishes in peer-reviewed journals? What proportion publishes in the content area in which they were trained or a closely related area?	descriptive statistics
▪ For each training award type, what proportion of trainees conducts independent research outside of NIH funding sources?	descriptive statistics
▪ How do these outcomes differ for trainees with clinical research training awards?	t-test
Research Training Progress How do NIH research trainees progress through NIH research training?	descriptive statistics
▪ What are the typical trainee profiles for NIH training mechanisms? For example, does a trainee typically move from the T32 to the mentored K to the R01, or does the trainee go directly from the T32 to the R01	descriptive statistics
▪ Does this profile differ for clinical research trainees?	t-test
▪ How do the trainees differ across the three Institutes in their use of NIH training mechanisms?	f-test
▪ To what extent do trainees use multiple NIH post-doctoral training awards, e.g., multiple T32s?	descriptive statistics
▪ Does age of trainee at time of training award affect the trainee profile?	regression
▪ Do other variables, such as trainee gender, affect the trainee profile?	descriptive statistics
Individual versus Institutional Awards Do NIH research training awards that require individual-level competition produce more positive outcomes than those involving only institutional-level competition?	
▪ How do T32 pre-doctorals and F31s who start their training awards at the same time compare in their career outcomes. ▪ How do T32 post-doctorals and F32s compare?	t-test
Comparisons among Institutes ▪ For all of the above, how do trainees in NIMH, NINDS, and NIDA differ?	f-test

RESOURCE AND COST ESTIMATES

The *core study*, as described in this chapter, involves the following tasks:

- Project management
- Finalize evaluation design
- Construct sample frame
- Draw sample
- Develop database
- Develop data collection protocols for each outcome
- Develop data collection instruments
- Develop data coding schemes
- Conduct Web and PubMed searches on 1,600 trainees
- Extract NIH grant and application data from NIH databases
- Establish inter-rater reliability for obtained outcomes and coding of outcomes
- Search NIH databases for NIH outcomes on 1,600 trainees
- Enter and code data
- Review and clean data
- Analyze data
- Report findings

A team comprised of the following members will be required to carry out these evaluation tasks:

- Project director with program evaluation experience
- Sampling statistician
- Research statistician
- Senior database developer
- Data administrator
- Web search and library staff
- Data entry staff

In *chapter 3* we estimated 4,800 hours and a cost of $350,000 to carry out the data collection and entry tasks alone using this core study approach, based on the sample estimate of approximately 1,600 research trainees. When the additional tasks required to complete the evaluation, as listed above, are included, we estimate that the entire core study will require approximately 6,200 hours, cost about $450,000 in total, and take about 18 months to complete. As noted in *chapter 3*, the supplemental data collection option of contacting the research training PI/mentor s would require approximately 500 additional hours and cost an estimated $36,000 additional.

References

Clouet, D.H. (1986). *The Career Achievements of NIH Predoctoraltoral Trainees and Fellows Supported by the National Institute on Drug Abuse.* Rockville, MD: National Institute on Drug Abuse.

Henderson, L., Lee, B., & Marion, A. (2001). Final Report on Three Focus Groups with Early Career Clinical Researchers about the K23 Award Program.

Joint Committee on Standards for Educational Evaluation. (2003). *The Student Evaluation Standards.* Thousand Oaks, CA: Corwin Press.

National Institute of General Medical Sciences. (1998). *The Careers and Professional Activities of Graduates of the NIGMS Medical Scientist Training Program.* Bethesda, MD: Author.

National Research Council. (2005). *Fostering the Independence of New Investigators in Biomedical Research.* Washington, DC: National Academy Press.

Pion, G.M. (2001). The Early Career Progress of NRSA Predoctoral Trainees and Fellows. Bethesda, MD: National Institutes of Health:

Appendices