06/05/03

A FEASIBILITY STUDY

OF

50 YEARS OF DNA: FROM DOUBLE HELIX TO HEALTH

June 2003

Prepared for the National Human Genome Research Institute

National Institutes of Health

By

Clare Von Secker, Ph.D.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

FEASIBILITY EVALUATION

EXECUTIVE SUMMARY

BACKGROUND

50 Years of DNA: From Double Helix to Health, a program sponsored by the National Human Genome Research Institute (NHGRI) in April 2003, commemorated three historic scientific events: the culmination of the sequencing of the human genome, the 50th anniversary of Watson and Crick's Nobel Prize winning description of the DNA double helix, and publication of a scientific report describing the future of the field of genomics and the role that the NHGRI and all of the National Institutes of Health (NIH) will play in enabling that future. The program consisted of a series of scientific, educational, cultural, and celebratory events conducted across the United States. Different events were designed to educate the public, to inform and stimulate scientific thought, to thank governmental leaders for their vital support of the NIH and of the Human Genome Project, or to inform national policy and opinion leaders about the success of science and its potential for further improving the nation's health.

EVALUATION PURPOSE

The purpose of this evaluation was to assess the feasibility of conducting an outcome evaluation of the educational components of *50 Years of DNA: From Double Helix to Health* that were aimed at high school teachers and students. NHGRI brought genomics to the classroom by developing on-line lesson plans, curriculum supplements, and other resources that could be used to teach about genetics and the ethical, legal, and social implications of genomic research during the celebration and throughout the school year and by encouraging teachers to participate in a national "DNA Day" on April 25, 2003.

An outcome evaluation would assess how well the educational components met program goals of increasing teachers' awareness and use of NHGRI web-based curricular and instructional resources; enhancing students' knowledge about and interest in science, genomics, the related ethical, legal, and social implications of genomic research, and scientific careers; and reaching out to minority and underserved school districts. The results of the outcome evaluation could be utilized by NHGRI to determine whether, and to what extent, education activities conducted as part of *50 Years of DNA: From Double Helix to Health* should translate to future NHGRI educational initiatives.

EVALUATION METHODOLOGY

Recommendations about the feasibility and design of an outcome evaluation of the educational components of *50 Years of DNA: From Double Helix to Health* were informed by (1) a review of the literature to determine the magnitudes of program effects estimated by other educational evaluations or research studies; (2) an assessment of the strengths and weaknesses of alternative methodologies and measures; (3) a review of the literature to identify the evaluation designs, data collection methods, and outcomes of programs with comparable goals; (4) analysis of the strengths and weaknesses of alternative data collection techniques; (5) synthesis of Office of Management and Budget (OMB) requirements for conducting federal outcome evaluations; and (6) examination of cost estimates of other federally funded outcome evaluations.

KEY EVALUATION QUESTIONS AND FINDINGS

1. At the conclusion of *50 Years of DNA: From Double Helix to Health*, will there be adequate justification to conduct an outcome evaluation?

An outcome evaluation of the educational components of *50 Years of DNA: From Double Helix to Health* is justified for measuring the extent to which program goals are met and for guiding decisions about future allocation of resources. Although the magnitudes of the program effects are expected to be small (Von Secker, 2000, 2002), evidence of program success can be detected by a well-designed outcome evaluation conducted with a sufficiently large sample. Results of the outcome evaluation can be utilized by NHGRI to answer questions about teachers' awareness of and satisfaction with the NHGRI web-based resources and to identify patterns of resource use that are most effective for attracting student interest, exciting and educating students about science and genomics, encouraging teachers to place greater emphasis on genetics and genomics in their classrooms, and serving disadvantaged students and schools.

2. If so, what are the most appropriate methodology and measures to use in evaluating the effects of the educational components?

The most appropriate methodology is a mixed methods design that combines both quantitative and qualitative measures to evaluate program effects. This approach can yield richer, more reliable, and more valid findings than one based on either qualitative or quantitative methodology alone. Qualitative findings offer contextual insights about the perceptions and reactions of website users that can be useful for refining objectives that have been stated in general terms (*e.g.*, "encourage students"). Quantitative findings provide statistical evidence to support inferences about the applicability of the findings to the population of all potential website users. The triangulation of information from different data sources is likely to yield the most comprehensive, credible, and useful information for helping NHGRI stakeholders make decisions about the strengths of the

educational components, areas for possible website modification, and strategies for improving teacher use of supplemental resources. Results of the outcome evaluation can be used to inform the NHGRI's broader policy deliberations related to increasing the public's awareness of science and the ethical, legal, and social implications of genomic research.

3. Are there comparable programs for which valid and reliable measures could be obtained to assess outcomes?

None of the programs provided examples of reliable or valid instruments that would be useful for measuring how well the proposed outcomes were met. New survey instruments and focus group protocols will have to be developed for measuring the effectiveness of web-based activities for students and teachers and the success of outreach efforts to traditionally underserved populations.

4. What would be the data collection strategies for an outcome evaluation?

The best data collection strategy would include a combination of quantitative and qualitative techniques including surveys, email correspondence, document studies of internal records and web logs, interviews with key informants from the NHGRI Office of the Director (OD), and focus groups of teachers and students who participated in activities associated with *50 Years of DNA: From Double Helix to Health* or not.

5. What would be the estimate of burden on NHGRI staff and the public for data collection for an outcome evaluation?

NHGRI will hire a contractor to conduct an outcome evaluation of the educational components of *50 Years of DNA: From Double Helix to Health*. The evaluation will determine whether and how teachers and students are using educational materials

available on the NHGRI website and the extent to which individual activities and resources met program goals. As a result of using an outside contractor to conduct the outcome evaluation, the burden on the NHGRI staff will be approximately four hours. The estimated burden on the public for completing instruments used to collect data for an outcome evaluation is 172 hours.

6. What would be the estimate of cost for an outcome evaluation?

The budgets for outcome evaluations of science education programs, particularly for those with a national focus, depend on the complexity of the evaluation questions and the level of detail with which they are explored. Outcome evaluation budgets in excess of $1,000,000 are not uncommon. The NIH Office of Science Education budgeted approximately $700,000 in 2001 to conduct an outcome evaluation of their first three science curriculum supplements, but is receiving additional funding of $175,000 from the National Science Foundation (NSF) to support the study. NSF awarded a grant of $1,600,757 to Columbia University in 2000 to conduct a national evaluation of the impact of summer professional development programs for science and mathematics teachers on student achievement. However some less comprehensive outcome evaluations of NIH educational activities, including one currently underway for NHGRI, have been conducted for as little as $50,000. NHGRI should budget a minimum of $50,000 to $100,000 to conduct the proposed outcome evaluation of the educational activities planned for *50 Years of DNA: From Double Helix to Health.* Costs of a focused outcome evaluation may be supported, at least in part, with funds available through the NIH One Percent Evaluation Set-Aside Program.

FEASIBILITY EVALUATION

BACKGROUND

50 Years of DNA: From Double Helix to Health, a program sponsored by the National Human Genome Research Institute (NHGRI) in April 2003, commemorated three historic scientific events: the culmination of the sequencing of the human genome, the 50th anniversary of Watson and Crick's Nobel Prize winning description of the DNA double helix, and publication of a scientific report describing the future of the field of genomics and the role that the NHGRI and all of the National Institutes of Health (NIH) will play in enabling that future. The program took the form of a series of scientific, educational, cultural, and celebratory events across the United States. Four goals of *50 Years of DNA: From Double Helix to Health* were to educate the public, to inform and stimulate scientific thought, to thank governmental leaders for their vital support of the NIH and of the Human Genome Project, and to inform national policy and opinion leaders about the success of science and its potential for further improving the nation's health.

Program events that were planned to educate the public, particularly high school teachers and students, are the focus of this evaluation. NHGRI brought genomics to the classroom by inviting teachers and students to join in the recognition of these historic achievements through direct mailings and frequent communication, by developing web-based lesson plans, activities, and curriculum supplements regarding the Human Genome Project, genomic science, and the nature of human genetic variation, and by encouraging teachers to participate in a national "DNA Day" on April 25, 2003.

The goals of the educational program components were to excite and educate students about science in general, and genomics in particular, as well as the related ethical, legal, and social implications of genomic research; to encourage teachers to explore genetics and genomics in the classroom in April 2003, as well as all year round; to raise teachers' awareness of free teaching tools available in genetics; to attract and encourage students, particularly minority students, to pursue scientific careers, especially in genomics; and to communicate with and create activities for minority students and teachers in disadvantaged school districts.

EVALUATION PURPOSE

The purpose of this evaluation was to assess the feasibility of conducting an outcome evaluation for the educational components of the celebration. An outcome evaluation would assess the extent to which program goals were met and guide decisions about further allocation of resources. Six study questions guided the feasibility study addressing the justification for and practicality of measuring whether program goals were achieved, namely:

1. At the conclusion of *50 Years of DNA: From Double Helix to Health*, will there be adequate justification to conduct an outcome evaluation?
2. If so, what are the most appropriate methodology and measures to use in evaluating the effects of the educational components?
3. Are there comparable programs for which valid and reliable measures could be obtained to assess outcomes?
4. What would be the data collection strategies for an outcome evaluation?

5. What would be the estimate of burden on NHGRI staff and the public for data collection for an outcome evaluation?
6. What would be the estimate of cost for an outcome evaluation?

EVALUATION METHODOLOGY

Recommendations about the feasibility of and optimal evaluation design for an outcome evaluation of the educational components of *50 Years of DNA: From Double Helix to Health* were informed by a review of the literature to determine the magnitudes of program effects estimated by other educational evaluations or research studies; an assessment of the strengths and weaknesses of alternative methodologies and measures; a review of the literature to identify the evaluation designs, data collection methods, and outcomes of programs with comparable goals; analysis of the strengths and weaknesses of alternative data collection techniques; synthesis of Office of Management and Budget (OMB) requirements for conducting federal outcome evaluations; examination of cost estimates of other federally funded outcome evaluations.

FINDINGS

1. Justifying an Outcome Evaluation of Web-based Instructional Interventions

This feasibility evaluation addressed the practicality of answering potential outcome evaluation questions about the effectiveness of the educational components of the NHGRI program *50 Years of DNA: From Double Helix to Health* aimed at teachers and students. The NHGRI website brought genomics to the classroom by posting an online curriculum supplement, *Human Genetic Variation*, about the basics of human genetics; lessons for teachers to accompany each of the components of an online multimedia unit *Exploring Our Molecular Selves*; independent classroom activities; a list

of volunteer mentors for classes; and links to other, free, high quality websites. Interested teachers and students can access NHGRI information resources quickly and at relatively low cost. There is an increasingly body of evidence to suggest that web-based resources are effective for educating students as well as generating interest in a topic (Chaparro & Halcomb, 1990; Guptill, 2000; Hargis, 2001; Marcoulides, 1990; Nulty, Halama, Dauzvardis, & Espiritu, 2000; Wilson & Harris, 2002; Winne, 1995; Worthington, Welsh, Archer, Mindes, & Forsyth, 1996; Zimmerman, Bonner, & Kovach, 1996).

The question of whether an outcome evaluation is practical can be quantified further by considering the magnitude of the effect anticipated by an intervention. Evaluation interventions typically are deemed "effective" when results are statistically significant (*i.e.*, $p < .05$). One of the limitations of this interpretation is that the calculation of the value of p for a test statistic depends in part on sample size and variability. Analysis of a very large sample may produce a statistically significant result that has limited practical value. Conversely, statistical tests conducted on small, highly variable samples can produce p-values that are not statistically significant even when the practical effects of a treatment are large.

Results of statistical analyses are more meaningfully interpreted when they are restated in terms of effect size (ES) estimates. ES estimates are standardized measures of the significance of statistical tests. The standardized measures allow comparison of outcomes with different metrics and yield results that are less sensitive to differences in sample size and variability. In educational research and evaluation, effect size values of .10, .30, and .50 are interpreted as small, medium, or large, respectively (Cohen, 1988). Effect sizes are useful for making decisions about whether the practical value of a desired

outcome justifies the expense of activities and initiatives designed to achieve that outcome.

The *Publication Manual of the American Psychological Association* (APA) encourages authors to provide effect size information in addition to results of statistical significance tests (American Psychological Association, 1994). But reviews of APA journals specifically (Kirk, 1996) and the literature in general (Keselman, et al., 1998; Thompson & Snyder, 1997, 1998; Vacha-Haase, 2001; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000) confirm that few evaluators and researchers heed that recommendation. Thus there is little empirical evidence to guide calculation of effect sizes particularly for curricular and instructional interventions. An outcome evaluation of the educational components of *50 Years of DNA: From Double Helix to Health* could contribute to the evaluation literature by providing guidance for estimating the magnitude of the effect of web-based educational activities.

An outcome evaluation of the NHGRI website and web-based activities is justified and practical for answering questions about the immediate effects of program components on participants' knowledge, attitudes, and career goals as well as for documenting whether, and in what ways, teacher utilize the supplemental resources. An outcome evaluation conducted with a sufficiently large sample could provide answers to these questions:

1. To what extent (effect size) do the challenging, independent activities on the NHGRI website attract student interest, and excite and educate students about science, genomics, and the related ethical, legal, and social implications of genomic research?

2. To what extent do materials and flyers mailed to teachers encourage them to explore genetics and genomics in the classroom all year round?
3. Are teachers aware of and using the free teaching tools available from NHGRI?
4. Do the educational components of *50 Years of DNA: From Double Helix to Health* attract and encourage students, particularly those in traditionally underrepresented populations, to pursue scientific careers, especially in genomics?
5. What activities targeted minority and underserved school districts and excited and educated students about science in general and genomics in particular as well as the related ethical, legal, and social implications of genomic research?
6. How can we utilize results of an outcome evaluation of *50 Years of DNA: From Double Helix to Health* to inform future NHGRI educational initiatives?

2. Recommended Evaluation Methodologies

Program evaluation approaches are differentiated by the philosophical assumptions that drive the kinds of questions the evaluation seeks to answer and the criteria used for making judgments about program effectiveness. The more quantitative approaches tend to focus on whether or not programs met specific objectives or goals for the target population. Decisions about program effectiveness are informed by statistical evidence that demonstrates whether the program has benefits for participants as a group; little attention is given to individual cases. More qualitative approaches aim to gain understanding or insight. Criteria for judging program effectiveness are more subjective and may vary from one person to another; there is no "right answer." The evaluation

approach influences the methodology or procedures for investigating the evaluation questions. Both approaches are valid and selection of one or the other depends on what the stakeholders will find most helpful when utilizing evaluation results.

Common Elements of Evaluation Designs. All evaluation designs, regardless of the evaluation approach, include three interdependent elements, namely the (1) sampling frame; (2) type and frequency of data collection; and (3) analytic model. These elements define who will participate in the evaluation, the burden of data collection on individuals and the population as a whole, and the evidence that will be used to make decisions about program worth, merit, and significance.

The Sampling Frame. The sampling frame describes the sample of individuals from whom data will be collected. Two types of samples are purposive, the most common in evaluation, and random. Purposive samples of cases or individuals are selected because they are likely to provide rich, in-depth information or because they are the only sample available given the practical constraints of the evaluation design (*i.e.*, a "convenience sample" of volunteers). Random samples of sites or program participants are best for estimating results for an average person and for making generalizations about program effectiveness for the population as a whole. A power analysis can be used to calculate mathematically the size of the sampling frame appropriate for detecting program effects. In practice, the selected sample size depends on the evaluation questions, the reliability of the data collection instruments, the planned analysis, the importance of the decisions made based on the findings, the need for credibility, the available resources and personnel, and the evaluation budget.

Two evaluations conducted for the NIH Office of Science Education found that the effects of curricular and instructional interventions are small (Von Secker 2000, 2002). When anticipated effects are small, the sampling frame should incorporate a larger sample size so that the evaluation design has sufficient power to uncover those effects (Cohen, 1988). That being the case, the recommended sample size for an outcome evaluation designed to answer questions about the use and effectiveness of the NHGRI educational materials is 1,000 teachers and students.

Type and frequency of data collection. The type and frequency of data collection are influenced to a large extent by the evaluation approach. The more structured, quantitative approaches rely on standardized assessments, surveys, structured interviews, and documents for data collection. Often data is collected once or twice at set intervals. The aim is to gather empirical evidence that will support valid generalizations about program effects. The more flexible, subjective, qualitative designs gather information using observational techniques, unstructured interviews, and focus groups. Data is collected frequently at formal and informal intervals. The aim is to gather evidence from testimonials and anecdotes that can be synthesized to generate a thick description that provides insight and illumination about program benefits for one or more persons.

Analytic models. The choice of analytic model is determined by the evaluation questions and data collection methods. The structured questions posed in evaluations with a quantitative orientation tend to be those that can be answered though application of descriptive and inferential statistical analysis, cost-benefit analysis, or cost-effectiveness analysis. Results based on statistical probabilities are used to make inferences about what is typical for a group. The more open-ended questions characteristic of qualitative

analysis eschew classical statistics and rely instead on techniques such as data coding and content analysis. Results based on qualitative synthesis are used to provide a thick description of a small number of cases.

Comparison of Six Designs. The consequences of decisions about the sampling frame, type and frequency of data collection, and analytic models become evident when one compares six designs commonly applied for evaluation of educational interventions: the case study; posttest-only; pretest/posttest; posttest-only with comparison group; pretest/posttest with comparison group; and longitudinal study. The first, the case study, is a qualitative approach that might be useful for gaining understanding about the unintended effects of programs on individual participants or for gaining better understanding of how the program works. The other five are classical quantitative designs that are useful for evaluating the extent to which program objectives were met. Any of these designs could be used to measure program outcomes; the final choice will best be determined by balancing theoretical considerations with practicalities such as the evaluation budget and available resources.

Case Study. A case study is a qualitative approach used for telling a story or illuminating understanding about one case or program that may or may not be typical of others. Data is collected frequently from a small, purposive sample. A team of evaluators is most likely to gather information using unstructured observations and interviews with key informants. Instead of relying on traditional statistical analytic models, this approach relies on the keenness of the perception of the evaluator, the consistency of agreement among the evaluation team regarding the findings, and synthesis of observational data into a thick description. Because each case is unique, findings are not useful for making

generalizations about program effects. However they can provide stories and anecdotes that extend understanding of quantitative findings and put a "face" on numeric results.

Posttest-Only. Posttest-only designs are one of the most common methods used to measure program effects. Often the sampling frame is a purposive sample of individuals who are either recruited for the program or who volunteer to participate. Standardized tests or surveys are collected from all program participants at the end of the program and results are analyzed statistically. Program effects are presented in terms of participants' average scores on some measure of interest after program completion. Lack of information about baseline characteristics of program participants limits the validity of inferences about causal relationships between program effects and observed outcomes. Findings may be criticized if alternative explanations could also account for measured outcomes. This design could be strengthened by testing a large, random sample.

Pretest/Posttest. The pretest/posttest design is the most prevalent type of repeated measures model used in program evaluation. This approach is an improvement over the posttest-only design because it takes into account the baseline or entry-level status of the program participants. Data is collected from all participants on two occasions, usually at the beginning and at the end of the program. Difference scores are calculated for each individual and statistically analyzed to determine whether the average amount of change in some measured outcome is significant. Inferences about program effects are stronger than those obtained with a posttest-only design because the pretest controls for pre-existing characteristics that could provide an alternative explanation for inferences about program effects. However, without a comparison group some critics may charge that improvements on the posttest were a result of insight gained by taking the pretest rather

than as a result of the program itself. Testing a large, random sample rather than a smaller or purposive one could provide evidence compelling enough to dispel this concern.

Postest-Only With Comparison Group. This design extends the posttest-only design by adding a comparison group to serve as a control. In both cases standardized tests or surveys are collected at the end of the program and results are analyzed statistically. However the sampling frame is expanded to include two groups, one that participates in the program and a matched comparison group that does not participate. Ideally, individuals in the sample are randomly assigned to one group or another to assure comparability. If random assignment is impossible, then preliminary analysis should be conducted prior to program implementation to assure that the relevant baseline characteristics of individuals in the comparison group match (are comparable to) those of program participants. Program effects are determined by comparing differences in outcomes for participants and non-participants. This robust design overcomes most of the theoretical limitations of the weaker posttest-only approach but the cost and logistical complexity of adding a comparison group constrain its use in practice.

Pretest/Posttest with Comparison Group. This repeated measures design compares program effects for two groups. Pretests and posttests, respectively, are administered prior to and upon completion of the program. Statistical analysis of the two groups investigates the significance of group difference on two test scores. Addition of a comparison group improves the validity of the pretest/posttest design because findings can be used to make causal inferences about program effects. This design is powerful for uncovering program effects but also costly and complex to implement. A simpler, less expensive posttest-only with comparison group design is a practical alternative for the

proposed NHGRI outcome evaluation as long as the evaluator verifies that the two groups are matched or controls statistically for pre-existing differences.

Longitudinal Design. Longitudinal designs are essential for measuring long-term program impact. Unlike posttest-only or repeated measures models, longitudinal designs measure program effects on at least three occasions – usually more – over a long period of time. Data may be collected prior to, during, and immediately after program implementation, but longitudinal designs are characterized by collection of follow-up data at one or more relevant intervals (*e.g.*, annually). Longitudinal designs are the only method for measuring whether observed program effects are sustainable. However this alternative design is impractical if the cost of follow-up is high or the anticipated program effect is short-lived.

Mixed Method Evaluations. The recommended methodological model for an outcome evaluation of *50 Years of DNA: From Double Helix to Health* is a mixed methods approach that combines the strength of a posttest-only with comparison groups design with the insight of a case study. Because a mixed method evaluation would combine qualitative and quantitative techniques, it can yield richer, more reliable, and more valid findings than one based on either the qualitative or quantitative methodology alone. Quantitative findings obtained from a random sample will provide statistical evidence to support inferences about the applicability of the findings for the population of all potential website users. Qualitative findings obtained from purposive samples will offer contextual insights about the perceptions and reactions of website users that and can be useful for refining objectives that have been stated in general terms (*e.g.*, "encourage

students to pursue scientific careers"). Thus the limitations of one approach will be offset by the strengths of the other.

The recommended sampling frame for the mixed method evaluation should include both a random sample of 1000 teachers and students who use the website and purposive samples of teachers and students selected because they are expected to provide detailed insights about why educational components were effective or not. Data can be collected electronically from a random sample of website users when they visit the website and from purposive samples of volunteers that provide feedback via email or through participation in focus groups. This mixed sampling will gather evidence that can be used to make generalizations about overall program effectiveness as well as to tell stories about outcomes for individuals or groups of individuals.

By using statistical and qualitative techniques to triangulate information from different data sources the analysis is likely to yield the most comprehensive, credible, and useful information for helping stakeholders make decisions about the strengths of the educational components, the areas for possible website modification, and what strategies would be best for improving teacher use of supplemental resources. It will also assist NHGRI in broader policy deliberations related to increasing the public's awareness of science and the ethical, legal, and social implications of genomic research.

3. Review of Evaluations of Comparable Programs

A broad search of the Educational Research Information Center (ERIC) database from 1975 to 2003 examined evaluations of programs whose goals or strategies were consistent with the educational components of *50 Years of DNA: From Double Helix to Health*. The valid and reliable measures that do exist are not appropriate for assessing the

planned outcomes on student excitement about and understanding of genomics or the related ethical, legal, and social implications of genomic research. The are no evaluations that describe how web-based science activities more successfully targeted minority and underserved school districts or were more effective at exciting and educating students from traditionally underrepresented populations. Table 1 provides a summary the goals, evaluation designs, and instrumentation used by four programs that can guide NHGRI's outcome evaluation plans.

Table 1
Summary of Four Program Evaluations

Program Type	Goals	Evaluation Design	Instruments	Implications
NIH Science Curriculum Supplements	Student interest and achievement; Influence teachers' instructional choices; Influence health-related behaviors	Posttest only with comparison group	Standardized assessments; work samples; anecdotal reports	Results support beliefs about the potential effectiveness of web-based curriculum supplements and other lessons
Teacher Professional Development (SWEPT)	Indirect effect on student interest and achievement; Change teachers' behaviors	Pretest/posttest with comparison group	Standardized tests; surveys; structured observations	Provides guidance for development of teacher surveys
Interactive Web Activities (K-Zone)	Engage and educate public	Case study	Observations and interviews	Demonstrates limitations of case studies
Instructional Video (HIV)	Educate the public; Influence attitudes and behavior	Pretest/posttest	Survey	Demonstrates short-term effectiveness of instructional video

NIH Science Curriculum Supplements. One of the educational components of *50 Years of DNA: From Double Helix to Health* is the on-line science curriculum

supplement *Human Genetic Variation*. This supplement is part of a series developed at the request of former NIH Director Harold Varmus, M.D., to support the goals of the National Science Education Standards[1]. The science curriculum supplements provide resources to help students understand a set of basic scientific principles, experience the process of inquiry, develop an enhanced understanding of the nature and methods of science, and recognize the role of science in society and the relationship between basic science and personal and public health.

In 2000 the NIH Office of Science Education (OSE) conducted a pilot evaluation of the first three supplements in the series, including the NHGRI supplement *Human Genetic Variation*[2]. A posttest-only with comparison group evaluation design examined outcomes for a sample of 17 matched pairs of biology teachers and their students who were randomly assigned to use the curriculum supplements or not. Data collection techniques included standardized assessments, samples of students' written work, and informal feedback from teachers. The instruments used in the pilot study are inappropriate for evaluation of NHGRI outcomes because they measure program impact on science achievement.

The pilot study provides anecdotal evidence to support the potential effectiveness of web-based curriculum supplements and other lessons. Virtually every teacher who used a curriculum supplement reported that the activities, particularly laboratory activities and games, stimulated student interest regardless of the level of their students. Teachers who used the supplements felt very strongly that in addition to being

[1] National Research Council. (1996). National science education standards. Washington, DC: National Academy Press.

[2] A copy of the pilot evaluation report and details about the costs of the evaluation are available from Dr. Bruce Fuchs, Director of the NIH OSE.

interesting, the activities in the curriculum supplements helped students apply creative and critical thinking skills to analyze the direct and indirect effects of scientific discoveries on their individual lives and on public health.

The NIH OSE, working with the National Science Foundation (NSF), currently is conducting a national study of the first three science curriculum supplements to build on findings from the pilot evaluation. The student assessments are the same as those used in the pilot study. Two additional valid standardized measures, The Test of Science Related Attitudes (TOSRA) and the 1999-2000 Local Systemic Change Classroom Observation Protocol developed for NSF could be useful in future NHGRI initiatives but do not address the specific questions of the proposed outcome evaluation. TOSRA is a 50-item standardized instrument that measures students' attitudes towards science. The 1999-2000 Local Systemic Change Classroom Observation Protocol is used to collect information about classroom implementation of curriculum supplements, lessons, and modules available on-line, and to measure implementation of national science reforms.

Teacher Professional Development Opportunities. Investment in teachers' professional development has been expected to be the most efficient way to effect improvements in student interest and achievement in mathematics, science and technology. One national initiative, the Scientific Work Experience Programs for Teachers (SWEPT), allows teachers to work in a variety of basic or applied research and development settings for two to eight weeks during the summer.

In 1998, Columbia University commenced the first quantitative impact evaluation of SWEPT.[3] A pretest/posttest with comparison group evaluation design examined

[3] The Principal Investigator for this study was Dr. Samuel Silverstein. He may be reached via email at Columbia University at this address: scs3@columbia.edu.

outcomes for a base year sample of 70 matched pairs of science and mathematics teachers who participated in SWEPT or not. The instruments used for data collection were standardized assessments in biology, chemistry, algebra, and geometry that were administered at the beginning and end of a course taught by a SWEPT or comparison teacher, pre- and post-teaching attitudinal surveys of SWEPT and comparison teachers and their students, demographic and educational background surveys of SWEPT and comparison teachers, demographic and socio-economic status surveys of students of SWEPT and comparison teachers, and surveys of the teaching materials and methods employed by SWEPT and comparison teachers. Individual items on the assessments and surveys were taken from the National Assessment of Educational Progress (NAEP) and other surveys available from the U.S. Department of Education's National Center for Education Statistics. While these items were used to answer questions different from those that are the focus of the proposed NHGRI outcome evaluation, they provide an example of secondary use of existing instrumentation. When the items used for national surveys match program goals they can be used by evaluators to develop a customized instrument at relatively low-cost.

Interactive Exhibit. The Committee on the Public Understanding of Science (COPUS) is a joint venture of the Royal Society, the British Association for the Advancement of Science and the Royal Institution that was organized for the purpose of raising the profile and level of public understanding of science activities in the UK. One program that addresses these goals is *The K-Zone*, a set of seven interactive exhibits on health topics (*e.g.*, cancer, heart disease). The aims of *The K-Zone* are to engage the public, to de-mystify science and health issues, and to effectively communicate the

science facts that will support responsible decision-making. The case study design of the impact of the seven exhibits was conducted at five contrasting locations. Pairs of evaluators observed and recorded interactions of a purposive sample of 269 individuals with the exhibits and interviewed 41 people who viewed the exhibits. Although evaluators gathered qualitative data about how people used and reacted to the exhibits they were not able to make any generalizations about the impact of *The K-Zone* on peoples' knowledge, attitudes, or health-related behaviors.

The evaluation, prepared by Evaluation Associates Ltd. for COPUS (1998), provides data that demonstrate the value of events designed to improve public understanding of science. The design also illustrates some typical shortcomings that limit the utility of program evaluation findings. Participants in the program were a self-selected group. Those who chose to respond may not have been a representative sample. A mixed method evaluation approach would have generated more useful information about website effectiveness. Surveys of a random sample of users could have provided quantitative information about whether and to what extent individual educational components were engaging, interesting, and educational. Web server logs that recorded patterns of user interactions could have provided qualitative information to support data obtained from interviews and illuminate understanding about why specific outcomes were observed or not.

Video-Based HIV Education Program. HIV infection is the fourth-leading cause of death in the world. While scientists and health educators understand the mode of transmission and methods of HIV/AIDS prevention, public misconceptions about this disease are widespread. Worldwide efforts to correct misconceptions and promote safe

sexual behaviors through HIV/AIDS education typically have been hampered by limited resources and a lack of well-trained personnel. Videotape designed to increase understanding about HIV/AIDS and to promote safe sex is a cost-effective alternative for improving the knowledge, attitudes, and HIV/AIDS prevention practices of teenagers and adults from varied intellectual and cultural backgrounds (Sawyer & Beck, 1991; Singh & Malaviyam, 1994; Stevenson, Gay, & Jasar, 1995; Stevenson & Davis, 1994; Torabi, Crowe, & Rhine, 2000).

Evaluations of the effectiveness of video education used pretest/posttest designs that measured changes in individuals' knowledge, attitudes, perceived susceptibility, or behaviors (*e.g.*, increase in condom use). While results of these evaluations show the immediate outcomes are statistically significant there is no evidence to support inferences that program impact is sustained. Future NHGRI decisions about the value of proposed education interventions should recognize that videotapes such as the one proposed for the *50 Years of DNA: From Double Helix to Health* celebration may be a valuable component of the educational program but may not have any lasting impact on participants' knowledge or attitudes about complex issues associated with genomics or the related ethical, legal, and social implications of genomic research.

4. Recommended Data Collection Strategies

Each data collection method has advantages and drawbacks given the evaluation's purpose, design, implementation, findings, conclusions, and utilization. Evaluation decisions are best informed by a mixed method of data collection that balances the limitations of one source of information with the strengths of another. Selection of data collection strategies is driven also by consideration of the extent to which collected

information will answer evaluation questions, the practicality of the proposed timeline, the cost-effectiveness of the methods, and the likelihood of revealing significant, but unanticipated, program outcomes.

Five realistic data collection strategies are pop-up web-based surveys, email correspondence, document studies, interviews with key informants, and focus groups.

Survey techniques.[4] Survey techniques can address quantitatively some of the same questions as those investigated with qualitative data collection methods. Properly constructed surveys are simple to administer and provide highly credible data in a cost-effective way. Further, a large number of responses are easily transformed for statistical analysis. However, surveys are limited in the extent to which they can tap into the contextual elements that explain why respondents answer as they do. Further, surveys do not allow for unexpected outcomes because individuals will only answer questions that evaluators ask. Complete reliance on self-reports (e.g., surveys or questionnaires) may not provide a complete picture of program effects.

Conversion of popular surveys such as TOSRA into electronic form is not a straightforward activity because the large number of items they contain and time required to collect responses preclude their suitability for the Internet (Supovitz, 1999). Long forms turned electronic can be excruciatingly slow to download, difficult to read on-screen, and generally unmanageable. The feasibility of conducting a web-based survey is improved if the designer develops a conceptual map that illustrates what data will be

[4] Website evaluation is possible with a generic NIH clearance. NIH's request for a generic clearance was written in broad language to ensure that all aspects of a website could be evaluated to ensure that intended audiences find the information provided on the Internet sites easy to access, clear, informative, and useful and to provide a means to better understand how to serve visitors to the NIH Internet sites. OLIB has created a repository of cleared surveys and is available to work with Institutes to develop OMB-compliant survey questions that will capture needed data.

collected and guides the process of survey development; creates an interface that is thematically consistent from page to page within the survey and with the website; divides the survey into manageable pieces but includes a link (*e.g.*, password or ID) for matching individuals' answers to different sections; uses pull-down menus for Likert-style responses rather than the traditional scales that stretch across a page; and minimizes large graphics and other speed or memory intensive routines so that users with less powerful modems download information as quickly as possible.

The proliferation of web-based survey software has created new options for data collection and analysis via the Internet. Many websites place a simple feedback banner on their sites that allow all users to select whether or not to provide feedback. A limitation of this sampling technique is that feedback from a convenience sample is likely to be biased; those who are very satisfied (or very dissatisfied) with the website are most likely to respond and results will not be representative of the target population. In addition, response rates for feedback banners tend to be less than one percent, a value that increases the length of time required for data collection and compromises the statistical validity of the results.

A better data collection approach is a pop-up survey administered to a random sample of website users. This method yields higher response rates, reduces the time necessary to collect feedback from a large sample, and controls the percent of visitors invited to complete responses. Pop-up surveys with one simple multiple choice question can be administered while users are waiting for website pages to be downloaded. A better method for the NHGRI evaluation may be to intercept visitors once they leave the website rather than when they log on. Alternative random pop-up surveys may include

other question types, including forced choice responses from a drop-down menu; multiple choice items that allow respondents to select one or more answers from a list of options; open-ended brief responses requesting an email address or other user information; or open-ended questions that ask the user to type in comments or reactions.

Commercial software is available from myriad sources to facilitate development and analysis of pop-up surveys. Web-based survey tools are economical alternatives to telephone or mailed surveys. Usually web-site users are directed to click on a link to a web page and fill out an evaluation or satisfaction survey. The web-based survey tool can collect, analyze, and display data. Web-based surveys can quickly and inexpensively be customized to create updates for policy-relevant subgroups of users.

Email Correspondence [5] Email is a simple, cost-effective means of collecting information about the extent to which independent activities on the NHGRI website attracted student and teacher interest and excited and educated them about science, genomics, and related ethical, legal, and social implications of genomic research. A purposive sample of site users who volunteer their email addresses could provide feedback in their own words about the short-term and long-term instructional impact of various web-site components. An advantage to gathering data via email is that anecdotal comments in the "users' voices" can be incorporated into reports evaluating consumer satisfaction with the website components.

Document studies. Documents studies involve analysis of a range of written or recorded materials that can include anything from public records such as newspaper archives, to internal records such as videotaped recordings of workshop presentations, to

[5] Email will be collected only from volunteers who choose to provide comments. Comments will be deleted once the data is analyzed by the evaluator. Permission to collect feedback via email should be incorporated with the OMB clearance to collect survey data.

personal diaries or letters. Document studies are conducted for the purpose of gaining insights about a program that cannot be observed or noted in another way. In general, document studies are a practical, cost-effective, and unobtrusiveness method of determining the historical trends of sequences of a program within the context in which they occur when sources are available and accurate. Document studies are not advised if the information is incomplete, inaccurate, questionable, lacks authenticity, or is difficult to access. Document studies will be useful for the proposed outcome evaluation because internal project records and web server logs can provide insights about program history, deeper understanding of the processes involved in program implementation, and records of participants' interactions with web-based program components.

Internal Records. Internal records include mission statements, strategic plans, logistics, budgets, manuals, correspondence, and descriptions of program development and modification over time. These internal records reflect institution's resources, values, processes, priorities, and concerns, and provide an unbiased record or history. These records are not prepared for the purpose of the outcome evaluation or at the request of the evaluator, but rather are a historical record of the evolution of science education initiatives and of the logistical considerations involved in program presentation.

Web Server Logs.[6] Web-based data collection is a promising methodology that allows researchers to study groups of people without compromising data quality (Schmidt, 1997). Analyses of web server logs patterns provide information not only about who uses a website but also how they use it, and can be utilized in evaluations of the effectiveness of instructional websites (Ingram, 1999). For example, logs can record not only the number of website visitors but also the number and type of activities that users

[6] Logs will include responses to pop-up surveys.

prefer, the amount of time users spend on the site overall and interacting with individual activities in particular, and the number of activities that are downloaded for use by teachers or students.

Data collected as part of the web logs can be saved in text files or databases. Text files are simpler records that record cross-sectional data that may or may not be displayed in various ways on the website. For example, text files can record the number of users that have accessed a site since a given date. Dynamic (database) systems, while more complex, have the advantage of allowing participants to provide longitudinal responses. For example teachers may volunteer to provide email responses that describe their expectations about the value of an educational activity or experience and then review their initial comments before they provide feedback about how well those expectations were met, how their understanding has changed, or what unanticipated outcomes they observed.

Interviews with Key Informants. Key informants are persons or group of persons with unique skills, perspectives, or backgrounds relevant for informing the program evaluation. In general, use of key informants is warranted if they are knowledgeable about the subtleties of program implementation, the needs of the program participants, or have expertise about information of interest to the evaluator. Use of key informants is less advantageous if the time required to select and get commitment is substantial, there is a risk of bias, or if the informants may influence the type of data obtained.

For NHGRI science education initiatives, members of the program staff can serve as key informants who provide "insider perspectives" about program implementation and success and "institutional memory" about the causes, rationale, and reasons for the

decisions and approaches that guided evolution of the program over time. Because the number of program staff who could serve as key informants is small, they can be interviewed individually. Advice and feedback from three or four key informants on the NHGRI staff will increase the credibility and utility of the outcome evaluation.

Focus Groups.[7] Focus groups combine elements of interviews, structured observation, and videotape data collection methodologies. Focus groups conducted by experts take place in a focus group facility that includes recording apparatus (audio or audiovisual) and an attached room with a one-way mirror for observation. There is an official recorder who may or may not be in the room. Focus groups were used initially as marketing research tools for investigating the appeal of various products. In that context, focus group participants were paid for attendance and provided with refreshments. As users outside the marketing arena (e.g., educators) have adopted focus group techniques these features have became less common.

Focus group sessions are composed of 8-12 people with shared characteristics relevant to the evaluation. Focus groups are distinct from discussion groups, problem-solving sessions, or decision-making groups in that they capitalize on group dynamics. Focus group techniques make explicit use of the group interaction to generate data and insights that would be unlikely to emerge without the interaction found in a group and allow firsthand insights into the respondents' behaviors, attitudes, language, and beliefs.

Focus groups are recommended over alternative methodologies such as individual interviews when the group interaction and dynamics are important for stimulating a richer response and to better capture the range of differences among individuals. The focus

[7] OMB clearance for focus groups is not required as long as fewer than ten participants are involved and the identical questions or script are not repeated with other groups of nine or fewer.

group process is valuable for challenging the clarity of individual communication and thinking and illuminating conflicting opinions. Focus groups are also desirable from a logistic vantage when staff resources, including the availability of qualified staff focus group facilitators to control and manage groups, support data collection from a large number of persons in a few groups but individual data collection with the same number of individuals would be impractical.

On-line Focus Groups. Online focus groups are a cost-effective and time-effective alternative to customary "face-to-face" focus groups that allow an invited group of eight to ten people to share comments for 90 minutes to two hours in a specialized chat room. Participants are able to view text, graphics, sounds, video or multimedia for evaluation and testing. As with traditional focus groups, the moderator prepares a series of questions in advance and focuses the discussion. Unlike face-to-face focus groups, complete transcripts of the session are available minutes after the conclusion of the session.

Focus groups are an important component of an outcome evaluation because data obtained from purposive samples of individuals who decide to use the web-based resources or not can illuminate interpretation of quantitative findings, recognize problems in program implementation; identify strengths and weaknesses of individual educational components, and generate awareness of unanticipated program outcomes.

Answering Evaluation Questions. A summary of the types of outcome evaluation questions that could be answered with each data collection method is provided in Table 2. Using multiple methods to answer each question will produce more comprehensive understanding of the impact of *50 Years of DNA: From Double Helix to Health*.

Table 2
Summary of Proposed Outcome Evaluation Questions and Data Collection Methods

	Evaluation Questions	Data Collection Methods
Educational Components Aimed at Teachers		
1	What are the demographic characteristics of teachers who were mailed flyers and free teaching materials?	Document Analysis Interviews with Key Informants
2	To what extent do materials and flyers mailed to teachers encourage them to explore genetics and genomics in the classroom in April 2003, as well as all year round?	Survey Email Focus Group
3	Are teachers aware of and using the free teaching tools available at the NHGRI website?	Survey Email Focus Group
4	What are the reasons teachers are or are not aware of and using the free teaching tools available from NHGRI?	Email Focus Group
5	What influences whether teachers explore genetics and genomics in the classroom throughout the school year?	Email Focus Group
Educational Components Aimed at All Students		
6	What activities were posted on the NHGRI website to attract student interest, and excite and educate students about science, genomics, and the related ethical, legal, and social implications of genomic research?	Document Analysis Interviews with Key Informants
7	What criteria were used to make activities on the NHGRI website attractive, interesting, and exciting to students?	Document Analysis Interviews with Key Informants
8	Do the challenging, independent activities on the NHGRI website attract student interest, and excite and educate students about science, genomics, and the related ethical, legal, and social implications of genomic research?	Survey Email Focus Group
9	What specific features of the NHGRI website attracted student interest, and excited and educated students about science, genomics, and related ethical, legal, and social implications of genomic research?	Email Focus Group
Educational Components Aimed at Minority and Underserved Students and School Districts		
10	Are teachers in minority and underserved school districts aware of and using the free teaching tools available from NHGRI?	Survey Document Analysis Focus Group
11	What activities targeted specifically minority and underserved school districts and encouraged students to pursue scientific careers, especially in genomics?	Document Analysis Interviews with Key Informants
12	What activities targeted specifically minority and underserved school districts to excite and educate students about science in general and genomics in particular as well as the related ethical, legal, and social implications of genomic research?	Document Analysis Interviews with Key Informants
13	How were activities that targeted minority and underserved school districts more exciting and educational than those aimed at majority students and school districts?	Email Focus Group
14	Do the NHGRI website activities attract and encourage students, particularly those in traditionally underrepresented populations, to pursue scientific careers, especially in genomics?	Survey Email Focus Group
15	What about the educational components of the *50 Years of DNA: From Double Helix to Health* attracted and encouraged students, particularly those in traditionally underrepresented populations, to pursue scientific careers, especially in genomics?	Email Focus Group

5. Estimate of the Burden on NHGRI Staff and the Public

NHGRI will hire a contractor to conduct an outcome evaluation of the educational components of *50 Years of DNA: From Double Helix to Health*. The evaluation will determine whether and how teachers and students, particularly those from traditionally underserved populations, are using the educational materials available on the NHGRI website and the extent to which specific features of the NHGRI website attracted teacher and student interest and excited and educated them about science, genomics, and the related ethical, legal, and social implications of genomic research.

Proposed data collection techniques include electronic correspondence, document studies, interviews with NHGRI staff, and focus groups. Pop-up electronic surveys, including invitations for email feedback, will be administered to a random sample of approximately 1,000 website visitors. The document studies will be conducted by the evaluator and will place no additional burden on NHGRI staff or the public. The evaluator will conduct interviews of up to one hour each with three or four NHGRI staff members who are able to provide insights about program development. Focus group protocols will be developed for six outcome-relevant subgroups of eight members each; each group will be asked different questions.

Estimates of the burden on NHGRI staff and the public for completing instruments used to collect data for an outcome evaluation are presented in Table 3. The estimated burden on NHGRI staff and the public for data collection is 176 hours. All of the costs, including those for the focus groups, should be included in the overall evaluation budget. Estimates provided in Table 3 can be included in clearance requests submitted to OMB.

Table 3 Estimate of Burden on NHGRI and The Public		
Item	Maximum Burden Hours	Estimated Costs
Electronic Survey of up to 10 minutes each administered to a random sample of approximately 1,000 website visitors. Some of these individuals will also volunteer to provide feedback via email.	100.0	$0.00
Document Studies	0.0	$0.00
Interviews with Key Informants	4.0	$0.00
Focus groups with six different target audiences: • 8 teachers who participated in DNA Day events in April 2003 X 2 hours X $20.00 an hour • 8 teachers who did not participate in DNA Day events in April 2003 X 2 hours X $20.00 an hour • 8 students who participated in DNA Day events in April 2003 X 2 hours X $20.00 an hour • 8 students who did not participate in DNA Day events in April 2003 X 2 hours X $20.00 an hour • 8 teachers at targeted minority and underserved schools X 2 hours X $20.00 an hour • 8 students at targeted minority and underserved schools X 2 hours X $20.00 an hour	72.0	$1,920.00
Total Estimated Burden and Costs	176.0	$1,920.00

Electronic Surveys. Any online survey or Web poll that can be accessed by the public must bear the OMB clearance number and expiration date signifying that the survey form has been reviewed by OMB and that it does not place "undue burden" on the public. The NIH has a blanket clearance from the Office of Management and Budget (OMB) to be able to perform user satisfaction surveys on all of its websites. The NIH request for a generic clearance was written in broad language to ensure that all aspects of a website could be evaluated and that the intended audiences find the information provided on the Internet sites easy to access, clear, informative, and useful and to provide a means to better understand how to serve visitors to the NIH Internet sites. OLIB has created a repository of cleared surveys and is available to work with Institutes to develop OMB-compliant survey questions that will capture needed data. A request for a clearance

may be submitted by email with the survey instrument to meadk@od.nih.gov. This expedited process takes 4 weeks from the time OLIB receives the request until it is cleared. A copy of the OMB clearance application form needed to conduct on-line surveys is provided in Appendix A.

In order to determine whether the NHGRI website activities meet the goal of reaching out to traditionally underrepresented populations of students and teachers in disadvantaged schools, the survey instrument will include a reasonable number of demographic questions that provide information about who is visiting the NHGRI website. The NIH OMB allows these questions as long as these questions are necessary to understanding the rest of the user satisfaction data being collected. While the number of questions considered "reasonable" is unspecified, the OMB submission must justify the need for each question by explaining how it contributes to evaluation of the website. For the proposed outcome evaluation the identification of respondents' race-ethnicity and school type (*e.g.*, urban) are warranted because this data will be analyzed to ascertain what policy-relevant subgroups of the target populations use the website activities and find them attractive and engaging.

Focus Groups. OMB regulations prevent asking the same questions of more than 9 members of the public (non-Federal employees) without appropriate clearance to do so. The planned analysis will conform to this regulation; students and teachers from majority and minority institutions will answer different questions about the effectiveness of the educational components of *50 Years of DNA: From Double Helix to Health*. Each proposed focus group will consist of eight members. No additional OMB clearance will be necessary if these conditions are met.

6. Estimated Cost of an Outcome Evaluation

An outcome evaluation could inform decisions about whether, and to what extent, education activities conducted as part of *50 Years of DNA: From Double Helix to Health* should translate to future NHGRI educational initiatives. An outcome evaluation would take into account the practical value and cost-effectiveness of the program activities by measuring not only whether the activities were developed and available but also the extent to which they accomplished goals of educating the public, promoting positive attitudes, and encouraging teachers to use on-line resources in their classrooms. An outcome evaluation would measure, for example, not only how many teachers looked at a module, but also how many used them afterward and their satisfaction with the product.

Budgets for outcome evaluations, particularly for programs with national rather than local audiences, often exceed $100,000; budgets in excess of $1,000,000 are not uncommon. For example, the NIH Office of Science Education is spending almost $1,000,000 to conduct an outcome evaluation of the effectiveness of three of their science curriculum supplements; 100 percent of this budget has been awarded to the contractor who will conduct all phases of the evaluation. The estimated cost of performing the SWEPT evaluation conducted by Columbia University was over $1,600,000; more than $1,200,000 of this budget was awarded to an outside contractor who developed data collection instruments, analyzed (but did not collect) the data, and prepared a final report. Both evaluations were similar in that they measured direct and indirect effects of program components on student achievement. The relatively high costs of site selection, instrumentation design, and data collection are to be expected for evaluations that are so comprehensive.

In 2001, the NHGRI Division of Extramural Research awarded a grant of approximately $50,000 to hire an organization to provide evaluation services to determine the impact of its multimedia educational kit titled *The Human Genome Project: Exploring our Molecular Selves*. This impact evaluation does not examine program effects related to achievement, but rather the extent to which the resource increased access to the latest information about the Human Genome Project, enhanced life sciences education, and facilitated presentations and discussions about the Human Genome Project, genomics, and genetics. The results of the evaluation will be utilized to make decisions about resource effectiveness, to encourage use of the resource by a broader audience, to update and improve the educational kit, and to help make decisions about producing similar resources.

Many of the anticipated outcomes of the educational activities planned for *50 Years of DNA: From Double Helix to Health* are parallel to those explored in the evaluation of *The Human Genome Project: Exploring our Molecular Selves*. A budget of $50,000 to $100,000 would be sufficient to determine whether, and how completely, NHGRI implemented the programmatic activities such as distributing material, designing web-based lessons and databases, producing videos, offering free on-line curriculum supplements, and targeting minority students and schools. The results of the evaluation can provide insights about barriers to implementation and can guide decisions about the value of similar, future efforts. A more extensive investigation of whether and in what ways, the educational activities on the NHGRI website attract student interest, encourage teachers to explore genetics and genomics in the classroom all year round, promote career interest among minority students, and reach traditionally underrepresented populations

and underserved school districts could easily increase the evaluation budget to $150,000 or more.

Costs of an evaluation may be supported, at least in part, with funds available through the NIH One Percent Evaluation Set-Aside Program. This program authorizes the Department of Health and Human Services (DHHS), under the Public Health Service (PHS) Act, to allocate up to one percent of appropriations for each PHS agency for the purpose of conducting evaluations of four primary types: needs assessment, feasibility study, focused process evaluation, or focused outcome evaluation. Although HHS identifies the amount of set-aside funds available to each PHS agency, the administration of the funds is the responsibility of the individual agencies. At NIH, the One Percent Evaluation Set-Aside is administered by the Office of Evaluation (OE), Office of Science Policy (OSP) within the Office of the Director (OD).

The NHGRI may use the findings of this feasibility evaluation to leverage this critical funding to defray costs of a focused outcome evaluation of *50 Years of DNA: From Double Helix to Health*. Applications for Requests for One Percent Evaluation Set-Aside funds via the Evaluation Express Award are available from the NIH OE. The requests must include identification of the project title and project officer; the primary purpose of the proposed evaluation; and a brief description of the program or activity under consideration that specifies the program goals, key questions to be answered, the study design, and a proposed budget.

REFERENCES

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th Ed.). Washington, DC: APA.

Chaparro, B. S., & Halcomb, C. G. (1990). The effects of computerized tutorial usage on course performance in general psychology. Journal of Computer-Based Instruction, 17, 141–146.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd Ed.). Hillsdale, NJ: Erlbaum.

Evaluation Associates Ltd. (1998). An evaluation of schemes and projects. Unpublished evaluation prepared for The Committee on the Public Understanding of Science. http://www.evaluation.co.uk/

Guptill, A. M. (2000). Using the Internet to improve student performance. Teaching Exceptional Children, 32(4), 16-20.

Hargis, J. (2001). Can students learn science using the Internet? Journal of Research on Computing in Education, 33(4), 475.

Ingram, A. L. (1999). Using web server logs in evaluating instructional websites. Journal of Educational Technology Systems, 28(2), 137-157.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey,M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. Review of Educational Research, 68, 350-386.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Marcoulides, G. A. (1990). Improving learner performance with computer based programs. Journal of Educational Computing Research, 6, 147–155.

McNulty, J. A., Halama, J., Dauzvardis, M. F., & Espiritu, B. (2000). Evaluation of Web-based computer-aided instruction in a basic science course. Academic Medicine, 75, 59–63.

Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. Behavior Research Methods, Instruments, & Computers, 29(2), 274–279.

Singh, Y. N., Malaviyam, A.N. (1994). Experience of HIV prevention interventions among female sex workers in Delhi, India. International Journal of STD AIDS, 5(1), 56-57.

Stevenson, H.C., Davis, G. (1994). Impact of culturally sensitive AIDS video education on the AIDS risk knowledge of African-American adolescents. AIDS Education and Prevention, 6(1), 40-52.

Stevenson, H. C., Gay, K.M., Jasar, L. (1995). Culturally sensitive AIDS education and perceived AIDS risk knowledge: Reaching the "Know-It-All" teenager. AIDS Education and Prevention, 7(2), 134-144.

Sawyer, R. G., Beck, K. H. (1991). Effects of videotapes on perceived susceptibility of HIV /AIDS among university freshmen. Health Values, 15(2), 31-40.

Supovitz, J. A. (1999). Surveying through cyberspace. American Journal of Evaluation, 20(2), 251–263.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the Journal of Experimental Education. Journal of Experimental Education, 66, 75-83.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent *JCD* research articles. Journal of Counseling and Development, 76, 436-441.

Torabi, M. R., Crowe, J. W., Rhine, S. (2000). Evaluation of HIV/AIDS education in Russia using a video approach. The Journal of School Health, 70(6), 229-233.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. Educational and Psychological Measurement, 6(2), 219-224.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. Theory & Psychology, 10, 413-425.

Wilson, S. P., & Harris, A. (2002). Evaluation of *The Psychology Place*: A web-based instructional tool for psychology courses. Teaching of Psychology, 29(2), 165-168.

Winne, P. H. (1995). Inherent details in self-regulated learning. Educational Psychologist, 30, 173-187.

Worthington, E. L., Jr., Welsh, J. A., Archer, C. R., Mindes, E. J.,& Forsyth, D. R. (1996). Computer-assisted instruction as a supplement to lectures in an introductory psychology class. Teaching of Psychology, 23, 175–181.

Von Secker, C. (2000). Pilot evaluation of the NIH science curriculum supplements. Report prepared for the National Institutes of Health Office of Science Education.

Von Secker, C. (2002). Effects of inquiry-based teacher practices on science excellence and equity. The Journal of Educational Research, 95(3), 151-160.

Zimmerman, B. J., Bonner, S., & Kovach, R. (1996). Developing self-regulated learners: Beyond achievement to self-efficacy. Washington, DC: American Psychological Association.