UNIVERSITY OF CALIFORNIA, SANTA CRUZ

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO NTA BARBARA • SANTA CRUZ

Silicon Valley – Educational
Research and Development Center

1156 High Street
Santa Cruz, CA 95064
(831) 459-3672
(831) 459-4618 FAX

26 July 2013

TO: Dr. Edith Gummer, Program Officer, National Science Foundation
Directorate for Education & Human Resources, Division of Research on Learning in Formal and Informal Settings

FROM: Rodney T. Ogawa, Professor and Principal Investigator

RE: Responses to Negotiation Questions for SMA: 1338411

Thank you for your note indicating that reviewers positively responded to the proposal we submitted to the SBE/EHR-BCC Program: "A Regional 'Datadrome': Developing a Comprehensive Regional Approach to Data Set Integration to Support Data-Intensive Research in Education in Silicon Valley." The following are the responses of our team to the 18 questions raised by reviewers and NSF staff regarding both intellectual and logistical issues. We have taken the liberty to re-order the questions so that they are now clustered by the following topics: alignment of the proposed project to NSF's mission; the specific activities to be undertaken to meet the project's objectives; the project's relationships with related at district, state, regional, and national levels; intellectual issues; and institutional and logistical matters.

A. Alignment with NSF Mission:

1. **From the NSF perspective, one of the most problematic aspects of the proposal was a lack of focus on science, technology, engineering and mathematics (STEM). How will you bring this focus to the forefront of your work?**

   STEM education will be a central focus in designing the regional education data set for Silicon Valley for two reasons. First, The Silicon Valley Educational R & D Center (SVERDC), which is one of the centers collaborating on this project, is presently focusing its work on enhancing learning in science and mathematics. For example, NLET (a partner in SVERDC and in this project) is completing an NSF-funded project to develop an agent-based model for 8th grade mathematics achievement, using data from San Jose Unified School District. In addition, SVERDC is partnering with a Silicon Valley start-up to provide middle and high school science and mathematics teachers with web-based instructional and assessment tools and to conduct research on the impact of these tools on student learning in science and mathematics. Second, it is widely documented that firms in Silicon Valley confront major challenges in recruiting employees who have adequate academic preparation in STEM disciplines.

Thus, in keeping with UCSC's institutional commitments and for Silicon Valley's regional strategic purposes, the proposed project will plan and design the regional education data set to enable researchers to answer questions about the causal vectors that explain the achievement of students in mathematics and science and the gap in achievement in science and mathematics that presently exists between students from low-income, racially, culturally, and linguistically diverse students and their more academically successful White and Asian peers. The project team will emphasize the inclusion of data that reflect learning outcomes in science and mathematics and data on factors both in- and out of school that prior research has linked to academic performance in science and mathematics.

B. Activities, Actors and Resources to Meet Project Objectives

2. The actual design of the work that will be conducted is underspecified in the proposal. Please provide more details about how the multiple objectives of the work will be achieved.

To design a regional database in Silicon Valley to support data-intensive research in education by integrating structured and unstructured data that capture both the depth of analytic levels and breadth of measures to integrate in-school and out-of-school factors affecting academic achievement, the project will attain 6 objectives (the 7th being to develop a regional model to inform national efforts) :

a. To establish a cross-disciplinary community of scholars, representatives of institutional agencies, and leaders of private companies to collaborate in developing a comprehensive regional data set that can be analyzed to understand the causal vectors of student achievement and the achievement gap.

A cross-disciplinary community of faculty at UC Santa Cruz, which includes education, psychology, economics, and computer science developed the proposal for this project and will provide expertise in leading efforts to meet objectives concerning the identification and mediation of issues regarding computation/analytics, technical infrastructure, and ethics/privacy.

NLET staff, with their presence in the public and private sectors of Silicon Valley, lead the work to identify and gain the collaboration of representatives of public agencies and private companies in the region that collect data that can contribute to the regional data set.

b. To identify existing structured databases.
c. *To identify existing unstructured databases.*

NLET staff and the PI, who partner in the Silicon Valley Educational R & D Center, will form the team that will conduct regional scan to identify prospective sources of data for inclusion in the regional data set. The SVERDC team will identify multiple sources within the region as well as relevant sources beyond the region (e.g., state) to construct a profile of data sources on the educational conditions, challenges and opportunities in the Silicon

Valley. The team will seek sources of both structured and unstructured data for in-school factors, ranging from data on instructional practices and student responses to measures of classroom and school conditions. The scan will also identify prospective sources of structured and unstructured data on non-school factors, focusing on regional stakeholders in housing, workforce development, transportation, health, mental health, and human services.

The SVERDC team will consult with public agencies identified as potential data sources to identify barriers to combining data from multiple sources in the comprehensive, regional data set. The potential barriers to data integration that are identified in this stage of the regional scan will be discussed with regional stakeholders at a project convening.

The team will scan the region's private sector to identify human and technological capital possessed by companies that can contribute to the development and analysis of the regional database. Many companies and some non-profits in Silicon Valley have developed products that can assist the project in the aggregation and linking of data from multiple and diverse sources, the application of advanced analytics to large-scale data, and data security.

While the SVERDC team is working to gain the participation of representatives of public agencies and private companies in the "community" to develop a regional education data set and conduct a regional data scan, UCSC researchers will be lead efforts to identify and address analytic/computational, technical infrastructure, and ethical/privacy issues and thus meet objectives 4, 5 and 6.

The SVERDC team will write a briefing paper that provides recommendations for the sources of structured and unstructured data from public agencies and private companies that could be included in the design of the regional data set. The team will also prepare a paper outlining the challenges that likely will be confronted in gaining access to data from these sources and recommendations regarding strategies for resolving these challenges.

d. *To determine the computational and analytic issues that must be resolved to make it possible to answer the core questions.*

A Professor of Psychology and Director of UC Santa Cruz's Center for Statistical Analysis in the Social Sciences will lead a team, which will include a quantitative researcher from education and an econometrician, in exploring issues that arise in analyzing a very large data set in terms of both sample size and number of variables. The team will consider the wide variety of statistical methods that might be employed to analyze these data, including the use of random coefficient (multilevel) statistical models to accommodate the hierarchical structure of the data in which students are nested within classrooms, classrooms are nested within schools and schools are nested within communities. The data analytics team will also address other issues that present analytic challenges, including that many student performance measures are categorical, the inclusion of repeated measurements for students, the potential for data on multiple members of students' families, and the analytic advantages that accrue from the availability of large samples. The computation and analytic team will prepare a paper that examines and makes recommendations regarding challenges and approaches to analyzing the comprehensive regional data set.

e. *To assess the technical, organizational and regulatory barriers to the aggregation and integration of those data bases into a single architecture. To establish the search, user and system analytics required to make such an architecture functional.*

Two members of the computer science faculty, who are affiliated with The Institute for Scalable Scientific Data Management, will lead work on this objective. This team of researchers will address five technical issues that arise in creating a large-scale educational database: ingestion, entity resolution, normalization, storage, and processing. The first three are standard database questions that arise when receiving data from multiple datasets with different information, data formats, and the like. The last two are big data questions involving the selection of appropriate infrastructure to support the storage, processing, and querying goals of the system---how best to store and process the data, with sufficient capacity to handle the expected data sizes and sufficient robustness to provide relatively high availability, and how to provide sufficient processing efficiency and capacity to support the expected analysis load. The team of researchers will draw on substantial bodies of technical expertise to explore how these issues will take shape given the particular characteristics of a large, complex education data set. This team will prepare a paper examining and making recommendations regarding the technical infrastructure that will can be employed to store and process the comprehensive regional data set to support data-intensive research in education.

f. *To examine the ethical, security and policy issues that inhibit the development of an aggregated database needed to answer the core questions.*

A professor of education who specializes in research ethics and is Director of The Center for Collaborative Research for an Equitable California and a member of the NLET staff, who has extensive experience working on issues of cyber-security, will lead the work to meet this objective. A data set of the breadth and depth that we are proposing presents many regulatory and ethical challenges to data integration and coordination. Drawing on expertise in the fields of the ethics of educational research, medical ethics, and trust communities and cyber-security, this team will identify and examine responses to the regulatory and ethical issues that will be confronted in gathering, organizing, storing, disseminating, and analyzing a large, complex education data set. This team will prepare a report that examines and outlines recommendations regarding issues of ethics and privacy associated with developing and analyzing the regional educational data set.

During the course of the project year, the PI, the Co-PIs who lead each of the teams focusing on a major area of work (Advanced Analytics, Technical Infrastructure, and Security/Privacy/Ethics), and NLET staff will form the project's Integration and Design Team. This team will meet monthly to share emerging results. These discussions will identify points of intersection across the areas of work and their implications for the overall design of the comprehensive, regional data set. In month 10 of the project, the Integration and Design Team will develop an overall set of recommendations for integrating the results of the work on each major area in a design for a comprehensive, regional data base, including approaches to analyzing the data set, the technical infrastructure to support the data set, and the security, privacy, and other ethical issues that must be addressed. The Integration and Design Team will draft a report, which will be distributed and discussed at the final convening of regional stakeholders.

Five regional convenings during the project-year will involve strategic Valley partnerships and education practitioners and researchers, community stakeholders and business and technology innovators. The first 2 convenings will be devoted to sharing and discussing results of the regional data scan to identify potential sources of data for inclusion in the regional education data set and the challenges that surface in discussions with representatives of public agencies and private companies who collect potentially relevant data. During convenings 3, 4, and 5, the project teams will report on the preliminary results of their examination of issues concerning advanced analytics, technical infrastructure requirements, and ethics/privacy, respectively. Regional representatives of public agencies and private companies will discuss these findings and work with research teams to develop recommendations for designing the regional education data set.

3. **Several reviewers expressed concern that the project is not doable in only one year. With such a short timeline, how will you ensure that the multiple objectives of the building collaborative capacity are achieved?**

   The feasibility of completing the proposed project by meeting its 6 objectives (as described in the response to the previous question) will be enhanced in at least two ways. First, the project will build on relationships that already exist among community members on the faculty at UC Santa Cruz and between NLET, the partnering private non-profit organization and representatives of public agencies and private sector organizations in the Silicon Valley region. As noted in the response to question 2, faculty members from education, the social, sciences, and engineering at UCSC who comprise the project teams already collaborated in writing the proposal for this project. And, 2 members of the NLET staff have developed working relations with administrators in various public agencies in Silicon Valley through their work to construct the data warehouse for San Jose Unified School District and another NLET staff member has relationships with high tech firms in Silicon Valley resulting from his work in early ventures in on-line education and IT based support for educational organizations. Second, 4 work teams will work simultaneously, with one team focusing on 2 project objectives and the other teams working to meet one objective each. The work of the 4 work teams will be coordinated and integrated by the integration team, which will be comprised of the leaders from each of the 4 work teams.

4. **The scope of work that will be conducted by the national Laboratory for Education Transformation is not clear. Please provide a more detailed description of who will be involved and exactly what they will be doing. In addition, $200 per hour appears an excessive funding level for this sort of work. Please provide a more reasonable estimate of support for this aspect of the project.**

   The funding requested for the consultation provided by the National Laboratory for Education Transformation (NLET) on the proposed project has been reduced to $125 per hour for 500 hours of work and a total allocation of $62,500. In consulting on the project, NLET will make the following important and extensive contributions. A revised budget and budget justification have been submitted to NSF.

   NLET will coordinate, manage, and facilitate five regional convening's designed to develop a model regional database in Silicon Valley. NLET will provide expertise in three primary organizational and functional capacities: 1) Pre convening planning, and logistics; 2) Convening execution; and, 3) Post convening follow up and next steps design.

Pre-convening planning will include coordination of advisory group and regional, state and national partners to include facilitation of pre-convening input with key players within the Silicon Valley, statewide and national network to include timing, calendaring and logistics, production and distribution of personal invitations and RSVP's to include the development, compilation and distribution of pre convening informational packets.

Convening Execution will include planning and organization of the physical and procedural convening requirements (including food, custodial, sound and visuals), plenary session design, focus group breakout design and logistics, participant agenda and feedback forms, facilitator agendas, overall meeting facilitation and the collection and distribution of minutes.

Post Convening responsibilities include the development and distribution of the participant data base, distribution of minutes for input and discussion following each convening, overall convening synthesis and summaries to guide follow up discussions and, in partnership with the Silicon Valley Research and Development Center, the design and development of follow up activities and next steps strategies fundamental to creating a model regional data base for data-intensive research in education.

5. The budget allowed for the five convening's, $10,000 per meeting, is excessive. Please provide a more reasonable estimation of the costs of these meetings.

UC Santa Cruz will provide facilities in its Silicon Valley Center to host the convenings. To cover the costs of food, custodial services, and participant travel, we are requesting that the amount for each convening be halved to $5,000. A revised budget and budget justification have been submitted to NSF.

6. The specific commitment from associated agencies needs to be strengthened. This may be an important outcome of the study, but the potential needs to be more strongly addressed up front. With whom do you have specific agreements to participate? Are these all of the necessary stakeholders? How will you ensure that you have the required agencies and individuals on board?

In project development, participating stakeholders were identified who were engaged in mutually beneficial partnerships and data-sharing agreements. Stakeholders were categorized in four primary work areas for collaborative research and analysis including Advanced Analytics, Technical Infrastructure, Regional Vectors and Security and Ethics, with a central Integration Team to coordinate and oversee the development of the regional data set. Within the categories identified in the Regional Vectors, city and county agencies, business and non-profits most impacted by the regional scan were initially approached for input and collaboration. Immediate specific agreements were established between Pierson Labs, Los Alamos National Laboratory, Resilient Networks, County Offices of Education, Project Cornerstone, Silicon Valley Community Foundation, and the Santa Clara County Department of Mental Health. Additional meetings have expanded the dialogue to include EBay, Google, Apple Computer, John W. Gardner Center Youth Data Archive (Stanford), U.C. Berkeley, CSU East Bay, San Jose State University, Santa Clara County Departments of

Social Services and Juvenile Justice, Santa Clara County School-Linked Services, United Way Silicon Valley, Digital Promise, Silicon Valley Leadership Group and Joint Venture Silicon Valley. Each of the listed organizations is responsible for developing, maintaining and/or responding to structured and unstructured multi-agency data resources. Each are also well positioned to participate and expand engagement as necessary to provide meaningful dialogue in identifying and addressing issues of cross-agency data sharing in support of whole student learning.

C. Relations To Related Efforts/Building a National Community

7. The proposal does not seem to sufficiently acknowledge other similar work that is being carried on by the states, with particular urban districts, and with the Regional Educational Laboratories, with their research partnerships. How might your project build on those endeavors and how is it different?

The project team has examined similar work at district, state, and regional levels. We intend to follow the progress of this work and to build on the experiences they accrue. For example, we have followed the work in Chicago to extend the school district's data system to the city, linking education data to data from the city's programs that offer services in health, social services, and the like. We have also contacted individuals associated with the work in North Carolina to accumulate a longitudinal education data set support research and policy development. We are also following projects in Regional Education Laboratories that employ large-scale data sets to improve education, including the Urban School Improvement Alliance of the REL Northeast and Islands, which is building the capacity of local education agencies to access and use data to improve school performance. We will draw on the work of these and other programs to inform the design of the Silicon Valley regional education data set along two major dimensions: (a) data collection and integration activities and (b) multi-agency and school district cooperative programs. While this work is relevant to the proposed project, the conceptualization of the Silicon Valley regional data set is distinctive in three important ways: Standards focus, Student focus, Big-Data focus.

Standards focus. The proposed project seeks to create a standards-based data system that unites, through common technology, protocols, standards and APIs, multiple agencies, organization, and school institutions. This project will consider how a common data standard, a common set of governance and regulatory standards, and a cooperative scaling and focusing standard can evolve from the process proposed. By analogy, what the team is proposing is similar to how previous standards (technical, data and organizational) have evolved and transformed commercial or government sectors outside of education. The effort is data standards plus organizational cooperation leads to a unified way conduct data-intensive research to support students and their families.

Student focus. Many efforts to integrate education data are not organized to link data to individual students, instead aggregating data at classroom, school, and district level. This presents challenges to researchers in teasing out potential causal vectors and mediating factors. It also challenges the ability of practitioners to employ data to differentiate instruction based on student need, strength, and interest. Finally, it does not allow students and families to have transparency into and agency with the aggregated and integrated data nor to participate directly in the benefits of multiple agencies working together.

Big data focus. "Big data" collection, integration and application is new to education. The proposed project will focus on bringing these capabilities to the combined world of social services, schools and students and doing so through the structure of this grant. The team has focused upon Silicon Valley in part to include the big data rich corporate and research efforts in the region. Conversations have already commenced with eBay, Google and SalesForces.com to understand how those systems work and can be helpful in informing the project's goals. Similarly, the involvement of the Baskin School of Engineering (UCSC) in this project is precisely because that engineering school has made its focus big data management. Similarly, the UCSC Division of Social Sciences has made its focus on statistical methods in service of social outcomes. Finally, the access to Los Alamos National Laboratory, a pioneer in big data and modeling analysis allows this team to explore bringing best of breed components form other sectors to the regional dialogue on improving the social and education fabric of the region.

8. **Multiple state and local agencies are attempting to do the sort of work that is laid out in this proposal. How will you connect to these other endeavors? How will you disseminate the efforts you are undertaking?**

As noted in the response to the previous question, the project team is seeking and following related work at district, state, and regional levels. In order to contribute to developing a national community that supports data-intensive research in education, the project team will identify eight types of entities and activities where data integration and cooperative agency activity are clustered. The team will collect state, regional and urban examples of this activity, will systematically review these efforts, interact with some of the organizations and programs, and will closely focus dissemination activities within these categories. They will provide guidance in the beginning to the project and, hopefully, the team's project will provide guidance to them during and at the conclusion of the project.
(a) state data integration efforts
(b) aggregation of education and student data
(c) federal programs and projects for data sharing and data access
(d) regional agency agreements for data sharing
(e) urban multi-agency and education cooperative programs
(f) dedicated non-profit organizations and programs
(g) dedicated university programs or centers
(h) research institute and REL's and partner activity
For example, members of the proposed project's team recently learned about a project recently funded by NSF, "Collaborative Research: Leveraging Matched Administrative Datasets to Improve Educational Practice and Long Run Life Outcomes: Toward Building a National Interdisciplinary Network," which seeks to develop a national network of scholars engaged in research employing large-scale, longitudinal data sets. Our project team will contact the Principal Investigators of this national project to learn about the emerging results

of their efforts and to link the proposed project to this national community.

Dissemination is often the weak portion of many grant-based projects, especially at the university research level. The involvement of NLET in the team is directed precisely at the framing, collaborating, naming, communicating, and involving of others in the work of the team, i.e, purposeful dissemination. The team will work with district, state, and regional organizations conducting related work to assist in solving the data "Tower of Babel" problem for which a big-data effort, not common to any of the existing efforts, can be a strong instrument in re-forming disparate research, policy, and practitioner activities. Through the Silicon Valley Education Research and Development Center, a joint center of UC Santa Cruz and NLET, this group will be involving community, regional and state leaders, policymakers and corporate stakeholders. The www.NLET.org website will be used for broad dissemination of web, data and video components.

9. **It was not clear to the reviewers how the findings of this study would be generalized to a wider national audience. How will the regional "datadrome " informs the development of a comprehensive national data set?**

   The U.S. has a long history of local and state control of education and various groups of education stakeholders have expressed concerns over the collection and connecting of data that can illuminate factors that affect educational outcomes. Thus, it will arguably require a fundamental institutional shift in order for a comprehensive national data set to be developed and utilized by researchers, policy makers, and practitioners. Research on the sociology of institutional change indicates that such change results from either technological advancements or social movements, or both. The regional "datadrome" that will be planned and designed by the proposed project potentially can contribute to promoting both technical advances and broad social networks, or communities, that combine to lead the nation towards compiling and employing a comprehensive education data set. As described in the response to questions 7 and 8, the proposed project is distinctive in its focus on standards, individual students, and the application of "Big Data" analytics. In planning and designing a regional data set with these foci, the project team will examine key issues, including the connecting of structured and unstructured data from a variety of public and private sources, computation and analytics, technical infrastructure, and ethics and privacy. This work will contribute to advancing the "technology" required to engage in data-intensive education research. The responses to questions 7 and 8 also describe how the project team will contribute to developing a national community to support data-intensive research in education. By linking to and sharing its results with other groups involved in developing and utilizing large-scale, longitudinal data sets at local district, state, regional, and national levels, the team that will plan and develop the regional "datadrome" for Silicon Valley will participate in a national network, or community, that can spur a social movement aimed at producing a comprehensive national data set.

10. **The advisory board seems to consist of mainly regional individuals. How might you expand the advisory board to get more national perspectives and expertise involved in the study? A connection to REL West might be a solution.**

The point raised by this question is well taken. The project team solicited advisory board members who can provide technical and institutional support for planning and developing a regional education data set. The project team will expand the board by 3 members to gain a more national perspective and draw on expertise from outside Silicon Valley. As suggested, we will connect with Regional Educational Laboratory West (WestEd), which can provide access both to its related programs and to the national network of RELs. In addition, the project team will connect with representatives of other organizations with a national presence and relevant expertise, including the Los Alamos National Laboratory and the University of Texas, which collaborated previously with NLET and UC Santa Cruz on an NSF-sponsored project to develop an agent-based model of 8th grade mathematics achievement using data from San Jose Unified School District. The project team has already received the support of the STRIVE Network, which has a national scope and success in developing inter-agency collaborations to advance education, and will pursue its representation on the advisory board.

D. Technical Issues

11. It was not clear to the reviewers exactly which variables or factors that influence the lack of success of students the project intended to address. While the exact nature of the factors will undoubtedly emerge in the work, please provide a discussion of how you intend to identify those factors as you operationalize the questions you intend to address.

We will take two general approaches to identifying the variables or factors that influence the lack of student academic success, generally, and academic success in science and mathematics, specifically. First, as we explain in the proposal, members of the project team from education and NLET will review research to identify factors, with an emphasis on malleable factors, that have been found to be associated with the academic performance of students. These factors will include in-school factors (e.g., teacher quality, instructional organization) and beyond-school factors (e.g., health, parenting practices) and will be applied to design elements of the regional education data set. Second, drawing from developments in the analysis of "big data," the project team that will examine computational and analytic issues associated with data-intensive research in education will consider the appropriateness of correlational analyses applied to big data for identifying in-school and beyond-school factors associated with student learning, with an emphasis on science and mathematics learning. To support this work, the project will draw on two sources: consultations with faculty in applied mathematics in UCSC's Baskin School of Engineering and private sector partners who are engaged in large-scale data analytics (e.g., Pierson Labs).

12. The language of the proposal focuses on the determination of causal vectors of the achievement gap; however, no randomized controlled design was indicated. How will the use of agent-based modeling of primarily administrative data provide such evidence of a causal nature?

Our "causal vectors" expression refers to underlying theoretical models. Many of the proposed analyses will use regression methods in non-experimental designs to predict a vector of dependent variables (y) from a vector of hypothesized independent variables (x), but with many potential confounding variables (that will be available in the proposed master database) also included in the model so that estimated path coefficients from x to y can tentatively be interpreted as causal relations. Furthermore, we anticipate researchers taking small random samples from the master database and conducting controlled experiments to confirm and clarify tentative causal relations detected in prior regression analyses. The use of small-sample experiments using a random sample from the master database could also take advantage of the new replication-extension design and analysis methodology described by Bonett (2012, *Current Directions in Psychology, 21,* 409-412) in which prior (objective) statistical information is incorporated into a new study so that causal treatment effects from small-sample experiments can be estimated with much greater precision.

13. How will the project incorporate the data that the state agencies collect? Will the project only be using the data that are developed locally that pass through to the different state agencies, such as student test scores, attendance and disciplinary behavior?

This project will link and aggregate current state educational agency data into coherent data structures to support the regional data network. Current state and federal requirements include annual reporting of demographic information; school safety; academic data; school completion; class size; teacher and staff information; postsecondary preparation; and fiscal and expenditure data. The project is designed to also identify and categorize regional structured and unstructured data sources currently not required collected or aggregated by local, state or federal agencies for accountability purposes, that are causal factors influencing student in-school and out of school achievement, preferences, activities, supports, and interventions. A stakeholder network flow map will show what the additional data resources look like in the region addressing current data silos and the resulting lack of integrated, cross-agency data analysis. Supported by both the practitioners and researchers of this project, advanced technologies and social science models will serve to define the requirements for integrating data from disparate, multi-agency data sources into a canonical, federated data-sharing model. Systemic divergence of how districts integrate legacy vs. new solutions will frame answers to the question, what aspect of a child is covered by what agency, what data? For example, what are the non-school causal vectors that explain school achievement for low-income, racially, culturally, and linguistically diverse students? How do these interact with school-based factors to explain achievement? How can analyses of all these causal factors improve policy and practice to enable stronger school outcomes and improved workforce readiness?

14. How will you know that you have been successful? What particular evaluation activities will you conduct and who will be responsible for them?

The project's Advisory Board will evaluate the proposed project. Board members will meet with the project team quarterly to evaluate the project's progress towards meeting its objectives and offer input for plans to move the work forward. At the end of the project, Board members will evaluate the extent to which the project served its purpose of planning and designing a comprehensive regional education data set for the Silicon Valley region to support data-intensive research in education by assessing the extent to which the project met its objectives: a) To establish a cross-disciplinary community of scholars, representatives of institutional agencies, and leaders of private companies to collaborate in developing a comprehensive regional data set that can be analyzed to understand the causal vectors of student achievement and the achievement gap; b) To design a regional-level pilot project to inform the development of a comprehensive national data set for data-intensive research in education; c) To identify existing structured databases; d) To identify existing unstructured databases; e) To determine the computational and analytic issues that must be resolved to make it possible to answer the core questions; f) To assess the technical, organizational and regulatory barriers to the aggregation and integration of those data bases into a single architecture. To establish the search, user and system analytics required to make such an architecture functional; g) To examine the ethical, security and policy issues that inhibit the development of an aggregated database needed to answer the core questions.

To conduct the summative evaluation, Board members will engage in the following activities: a) review notes recorded for all convenings of project stakeholders to determine the representation and nature of participation of scholars from across disciplines and the representation and nature of participation of representatives of public agencies and private companies in the region; b) review papers prepared by work groups on the following issues: identify public and private sources of structured and unstructured data for inclusion in data set, examine computational and analytic issues, examine technical infrastructure, and examine ethical and privacy issues. Also read the final report that integrates the work of these teams and recommends design parameters for the regional education data set; c) interview participating researchers and agency and company representatives.

E. Institutional and Logistical Issues

15. You indicate that there does not need to be a data management plan, yet the proposal is dependent upon the aggregation of data from multiple sources. While none of that data might be held by any of the individuals, the security and management of that data still needs to be addressed. Even if the data remain in the control of the agencies that have collected them, please address management issues.

The purpose of this project not to aggregate data from multiple public agencies and private enterprises. Rather, its purpose is to plan and design a comprehensive regional data set in Silicon Valley that will support data-intensive research in education. Consequently, the project will have no data to manage.

Indeed, in planning and designing the regional data set, the project team must develop a plan to manage data aggregation, storage, access, and the attendant issue of security. When the plan and design for the regional data set is implemented, then the data management plan will be in place and enacted.

16. The proposal does not indicate that Human Subjects permission has been granted for this work. No work may move forward that addresses any human subject data until either an exemption or approval has been received from the requisite IRB. Please address this issue.

Again, the purpose of the proposed project is not to collect, aggregate, or analyze education data. Rather, its purpose is to plan and design a comprehensive regional data set that will support data-intensive educational research. Consequently, there will be no "human subjects" during this phase of the work. In designing the regional data set, one of the project teams will be devoted to examining issues of ethics and privacy, which will have direct implications for the handling of human subjects issues when the data set is aggregated and ultimately analyzed.

Due to current concerns over data security and privacy that have arisen in multiple public domains, including education and national security, the PI will submit a request to the UC Santa Cruz's Institutional Review Board to receive an exemption prior to commencing the proposed project.

17. **The project does not mention any evidence of prior support from NSF funding. While the exact project probably does not have such support, a number of the co-PIs have been supported by NSF in their development of the expertise that will be brought to bear on the work of the proposal.**

    This project has not previously received support from NSF funding. However, co-PIs have been supported in work that contributed to the development of the expertise that they will bring to the proposed work. Douglas Bonett, who will lead the team examining analytic and computational issues regarding large-scale, complex education data sets, received funding from NSF in 2003 for work that involved new kurtosis estimating methods that lead to improved confidence intervals and tests for measures of variability. This NSF funded research could relate to the proposed project because the new kurtosis estimates also improve tests and confidence intervals for path coefficients in the types of covariance structure models that we could fit from the data collected in the proposed project. Eduardo Mosqueda is co-PI on a project that is supported by NSF funding and studies the impact of professional development on the instructional practice of educators who teach science to English learners. This work is could relate to the proposed project by providing a Key Personnel with expertise in assessing the impact of teacher characteristics and practice on science outcomes for students. Consultants from NLET—Gordon Freedman, Marcy Lauck, and Bill Erlendson—are involved in a study supported by NSF funding that to develop an agent based model for 8th grade mathematics achievement, using data from a district in Silicon Valley. This work will have relevance to the proposed project by enabling these team members to develop expertise in the analyzing large-scale education data sets.

18. **At least one of the co-PIs appears to have more than 2 months salary from NSF. This needs to be clearly explained in the budget justification for each instance.**

    The PI on this project was listed as a Key Senior Personnel on a proposal pending with NSF and budgeted to receive 2 months of salary. However, that proposal was not funded.

    Eduardo Mosqueda is included as a Key Senior Personnel on this project and is budgeted to receive .5 month of salary. However, Dr. Mosqueda is co-PI on 2 projects funded by NSF and is budgeted for a total of 3 months of salary. Dr. Mosqueda will serve as a Key Senior

Personnel without salary for the work he contributes to planning and designing a regional education data set for Silicon Valley.