

# Modeling the Outcomes of the Bridges to the Baccalaureate Program: Feasibility Assessment

August, 2003

## **Background**

Each Bridges to the Baccalaureate program supported by NIGMS is composed of several component parts and it is difficult to say with certainty what the impact has been of each of these various interventions (i.e., to identify the key “active ingredients” of the programs). Although NIGMS could collect more data on the programs and their participants, it is not clear whether methods exist to adequately quantify the independent effect of these individual components on success of the participating students. So, prior to launching large-scale data collection to address this question, NIGMS conducted a “feasibility assessment” by asking several experts in multivariate statistics and evaluation for advice on the availability of methods to generate meaningful results given the nature of the program, the outcomes of interest and potential influences on them, and the amount and quality of data available for the analysis.

Independent assessments were solicited from:

Dr. David Cordray, Professor, Department of Psychology and Human Development, Vanderbilt University; an expert in the use of multivariate analysis in evaluation and Past-President of the American Evaluation Association (AEA)

Dr. Lee Sechrest, Professor Emeritus, Department of Psychology, University of Arizona; an expert in the development of methods for research and data analysis in field settings and another Past-President of the AEA

Dr. Charles Sing, Professor of Human Genetics, University of Michigan; an expert in the development and application of multivariate statistical approaches to study complex genotype-phenotype relationships

## **The Feasibility of Multivariate Approaches to Determine “What Works”**

In his assessment, Dr. Cordray emphasized that while, *in theory*, it is feasible to derive estimates of the impact of Bridges program components on student outcomes using existing evaluation methods, in practice it will be difficult and feasible only under a stringent set of conditions. In particular, regression-based models can be used to estimate the effects of components *if* reliable and valid indicators of all key variables (i.e., which program components each student was exposed to, the degree of exposure, participant characteristics, institutional characteristics, and outcomes) are used. Importantly, for each variable included in the model, the reliability and validity of the measures must be equivalent; otherwise the size of the coefficients will be inaccurate. Empirically, it will be difficult to satisfy these conditions. In addition, it is expected that program

components will be interrelated. The interpretability of coefficients of effects will depend on whether components are conceptualized as having effects on outcomes that are direct, indirect or both. A program model of the interrelationships among components must be specified.

Dr. Sechrest notes that it may not be possible to make strong causal inferences from any evaluation that is designed. This is because individual Bridges programs differ in the quality of the students and in the physical and human resources available to invest in them. These are confounding variables that will lead to misleading conclusions about what are the effective program elements. Furthermore, introduction of a comparison group is viewed as almost a necessity for making strong causal attributions of outcomes to elements of the Bridges program. Dr. Sechrest also raises issues of measurement and data quality. He concludes that the data currently available are not sufficient for the proposed evaluation. Furthermore, he describes the difficulties in measuring the variables needed in any quantitative model of outcomes. These variables include characteristics of the participants, institutions, and indicators of the components of the programs in which each student participated. In order for participation in a program component to be a useful predictor of program outcome, the component must be defined in the same manner across grantees. If it is not, differences among programs that include that component will “swamp out” any real effects. Even if positive effects of a program component were to be detected, one would not know exactly how to advise programs to implement that component. Finally, in the absence of sound theoretical models of how the Bridges programs produce their effects, statistical approaches are likely to produce disappointing results. Large-scale data collection and analysis involving a broad spectrum of Bridges grantees is not recommended.

Dr. Sing focuses directly on the question of the availability of statistical methods for making causal inferences, and their applicability to the Bridges programs. He notes that there are no statistical methods to deal with an unsupervised and unvalidated data collection scheme such as the one represented by the E-STAR database. The correlations between predictors inherent in such data prevent the estimation of the independent effects of a causal component. The resultant confounding of the effects of predictor variables is also a consequence of the uniqueness of each Bridges program. Unless one reduces the number of predictors considered, adopts a rigorous data collection scheme, and assumes that the effects of the other predictors are randomly associated with the selected predictors, it is unlikely that the independent effect of any particular predictor variable can be estimated or that the conditional effect, estimated using partial regression-like techniques, would be generally applicable to Bridges programs. Furthermore, any conclusions about the impact of a particular component will be highly dependent on the context defined by these program characteristics. General inferences having wide applicability will be difficult and inappropriate because such generalizations may be unrepresentative of any particular Bridges program.

## Summary and Recommendations

All three experts emphasized that, for the results of any evaluation to be meaningful, it is essential that the question(s) to be addressed by the evaluation is (are) identified clearly prior to initiating data collection. Furthermore, data collection must be much more systematic, intensive, and more thoroughly validated than currently possible using data collection mechanisms such as E-STAR. Data of the quality needed to conduct the proposed modeling study are the result of collection processes with a high level of organization, oversight, and cooperation. In view of the need for detailed and high-quality data, none of the three experts endorses the idea of a large-scale evaluation effort designed to address general questions regarding what strategies work among the Bridges programs.

As noted by Dr. Sechrest, the weakness of any evaluation will be the inability to make causal inferences, particularly in the absence of strong theoretical models of student outcomes. To be interpretable, the models of student outcomes must include all of the major determinants of outcomes. At the same time, however, sample sizes will be adequate only if the models are not too complex. Measuring the variables to be included in the model present many problems, with program elements being particularly difficult to characterize. A program may be different for each participant, leading to the very difficult task of saying what “the program” is and how it might be described. Considerable effort would have to go into characterizing, in measurable terms, what exactly the Bridges “treatment” is, the mechanisms by which the various components of the Bridges program are expected to have an effect on student outcomes, and identifying and measuring the other determinants of outcomes so that they may be included in the model as statistical controls. Sample sizes at some of the larger Bridges institutions may be sufficient to build a model if it is not too complex, students are homogenous, there is variation in the type of experiences each student has in the program, and the individual experiences of each participant can be measured reliably. To yield interpretable results, all measures included in the model need to be equally reliable and valid.

The experts were asked to identify what types of evaluation they thought might be feasible. Their reports suggest that while it may be possible, in theory, to derive an index of the effect of program components within Bridges, this will be successful only if a stringent set of conditions can be met. Assuming these difficult theoretical, measurement, and management issues could be solved adequately, useful information could be gained from evaluations that are focused on specific Bridges programs (or perhaps subsets of programs) and designed to address specific questions tailored to the implementation and goals of the individual programs under study. This approach is consistent with the current NIGMS policy of encouraging grantees’ evaluation of their own programs. As noted by Dr. Sing, it should be recognized that any conclusions about the impact of a particular component from such a study would be highly dependent on the context defined by the program characteristics. It would be difficult to make general inferences having wide applicability.