Draft 12/7/06

Plan for Analyzing ECLS-K Data in Order to Inform the 2007 Pilot Study[1]

The main purpose of re-analyzing the ECLS-K data is to determine relative costs, accuracy, reliability, and validity of SES variables selected and constructed in several ways from various combinations of census areal units and population subgroups. The major focus will be to compare:

a) Relationships among academic assessments,[2] parent self-reports, and aggregate SES estimates obtained from larger areas, e.g., ZIP code tabulation areas (ZCTAs), from subsets of households matched to the student on specific demographic characteristics (e.g. race and age).[3]

b) Relationships among academic assessments, parent self-reports, and aggregate SES estimates obtained from smaller areas (e.g. block groups or tracts) but from households matched to students with or without matching on demographic characteristics.

If we use area means for all selected households in an areal unit that meet defined criteria, e.g., containing a youth of appropriate age and race, then the number of selections will vary across areas and the variances of the means will vary correspondingly. (What will be the consequences of this variation in the reliability of aggregate estimates?) If we sample a fixed number of households per areal unit, there will be no such variation, excepting the effect of sampling from a finite population and observable differences in heterogeneity between area-subpopulation combinations. But to eliminate differences in variability due to sample size, we will either have to accept a very small number of households for each population subgroup or to draw from areas of varying size, depending on the availability of matching cases. For example, if a census tract contains very few white 9-year olds, one might have to go to a larger area to obtain relevant households for that subgroup.

c) Model-based[4] relationships among academic assessments, parent self-reports and SES estimates obtained from larger areas, e.g. ZCTAs, but from a small number[5] of households matched to the student on specific demographic characteristics (e.g. race and age).

[1] First draft by Frank Jenkins, WESTAT, 12/4/06. Revision by Robert Hauser and Sal Rivas, University of Wisconsin-Madison, 12/07/06.

[2] See http://nces.ed.gov/ecls/KinderAssessments.asp.

[3] A secondary analytic issue here and in analyses based on smaller areas (as in section b immediately below) is the tradeoff between greater specificity in the population used to create the aggregate and the smaller number of relevant households meeting restrictive population definitions. That is, are we better off using characteristics of entire areas, as in Beveridge's work, or using characteristics of subpopulations?

[4] By "model-based," we mean that characteristics of randomly selected households are treated as error-ridden indicators of the true characteristics of the student household.

[5] By "a small number," we mean no more than four or five, and possibly as few as two or three households.

d) Model-based relationships among academic assessments, parent self-reports and SES estimates obtained from smaller areas (e.g. block groups or tracts) but from a small number of households selected at random, with or without matching on demographic characteristics.[6]

The model here is as follows:

$$X_{ijk} = \xi_{jk} + \varepsilon_{ijk} \qquad (1)$$

where X_{ijk} is the value of the k^{th} SES variable, e.g., household income, from the i^{th} random draw from a subpopulation relevant to the j^{th} student; ξ_{jk} is the (latent or true) value of the k^{th} SES variable in the j^{th} subpopulation, and ε_{ijk} is an independent, random error. However, because households are the units drawn at random, we do not assume that Cov(ε_{ijk} , $\varepsilon_{ijk'}$) = 0. That is, there may be correlations between the errors in different socioeconomic variables drawn from the same household, conditional on the relationships among unobservables. Note that the units here are subpopulations within areas, thus, the covariances of the unobservables depend both on between- and within-area variation.

The problem with model-based estimation here is that it treats the true SES variables as unobservables, raising the question whether it is practical to use any such scheme in NAEP reporting. If we are stuck with observables, then the error-ridden aggregate estimates will have to do. That will force us to higher level units, possibly with odd variations in reliability across units, depending on the selection scheme for households within units. However, in these preliminary analyses, we think it is important to compare findings based on this model with those based on aggregate socioeconomic characteristics.

For at least two reasons, we need to look at relationships involving academic achievement as well as relationships between socioeconomic variables in the ECLS-K and those from Census data in these analyses. First, academic achievement is the *only* criterion available in the 4th and 8th grade NAEP samples in which we will also have data from the enhanced background questionnaire (EBQ). That is, in the main 4th and 8th grade 2007 pilot samples there is no *socioeconomic* criterion, e.g., a parent report, for the validity of EBQ measures. In the 8th grade ECLS-K sample, we will have parent reports as well as NAEP achievement measures, but no EBQ items. Because parent reports and EBQ items never occur in the same sample of students, there is no direct way to tell whether use of the EBQ measures would lead to greater correspondence between a socioeconomic index based on student reports (with or without supplementation by Census data) and a socioeconomic index based on parent reports. Academic achievement is the only available criterion for a direct comparison of validity between indexes with

[6] Our initial assumption is that there will be sufficient numbers of cases in ZCTAs to draw random matches for each student from appropriate subsets of households, but that there may not be enough cases to do this in smaller areas. However, this is an empirical matter to be explored when the microdata become available.

and without EBQ measures. However, using academic achievement in 2007 NAEP as a criterion, we can compare the relationships between socioeconomic indexes that included and exclude EBQ items as observed in the main 8th grade NAEP pilot sample and in the 8th grade ECLS-K sample.

Second, we won't know which socioeconomic characteristics of households are the best candidates for a new measure unless we know the relationships of those characteristics, as reported by parents in ECLS-K, with academic achievement. For example, should parental occupations be characterized in broad Census-like categories? By levels of occupational prestige? By typical levels of the education of jobholders? By typical levels of income of jobholders? Should our interest focus on obtaining a sound measure of household income or of mother's educational attainment? The issue here is not to decide what SES is, but rather to choose a minimally adequate set of valid, reliable, and easy-to-measure variables from the larger set of socioeconomic variables potentially available to us in future rounds of NAEP.[7]

Is there a basis for choosing among the array of candidate SES variables that does not rely in part on their relationships with academic achievement? Consider, for example, the correlation between a candidate socioeconomic variable, either from Census data or the NAEP background questionnaire, and a corresponding parent report. Does the size of that correlation tell us whether the construct in question represents a household-level socioeconomic resource for academic success? That is, would we want to choose that subset of socioeconomic variables from the Census data of NAEP background questionnaire that are most highly correlated with parent reports? What if we learn from the ECLS-K data that one construct, say, household income, which students cannot report accurately (and which we cannot, in any case, ascertain directly in NAEP) is the most important socioeconomic resource for academic achievement? Should we want to drop it merely because students can't tell us about it? If income proves exceptionally important as a condition of academic success in ECLS-K, we think that the appropriate inference is that we should look for the best possible proxy measures of income to use in NAEP.

There is a trade-off between area size and specificity of the matching, as smaller areas will not have enough matching households for a fine-grain match with students. This will affect both strategies for selecting data from areal units.

Another purpose of the ECLS-K reanalysis is to determine if area means or randomly selected households yield the most accurate and valid estimates of relationships of student's SES characteristics with academic assessments, taking into account what we can learn about relationships between those areal data and parent reports of SES.

The plan is to use the ECLS-K data in the following way:

[7] We can cross-validate inferences from these analyses, first looking at the achievement measures in ECLS-K before NAEP data become available and then looking at the NAEP achievement measures in the subset of ECLS-K cases included in 2007 NAEP.

a) We will start with the dataset created by Beveridge for his previous ECLS-K study. This dataset has the assessed student files matched to geographic area through geocoding.[8] Aggregate SES data from the 2000 Census long form is appended to each case. Each case also has the students' parent questionnaire data appended. *Aggregate* SES data are of limited direct interest here because we already have Beveridge's analyses based on these data. However, as noted above, findings based on aggregate characteristics of entire tracts or ZCTAs are a relevant point of comparison for the analyses proposed here.

b) The dataset will be augmented by matching student records to the long-form data from Census 2000 for randomly chosen individual households and/or for aggregates of households in their geographic area for each of 3 levels of geography (block group, tract, and ZCTA). This will permit analyses in which all SES assignments are based on a single level of geography and, also, analyses in which varying levels of geography are combined, depending on the prevalence of defined demographic subgroups in a locality. Another important issue to address here is rates of success in geocoding at the block, tract, and ZCTA levels, which may vary by geographic area, e.g., between urban and rural areas.

c) The combination of 2 types of SES estimates (means and randomly selected households) and 3 levels of geography (block group, tract and zipcode) will result in at least 6 sets of data to be analyzed. There will be additional sets of data to be analyzed if we vary the number of randomly selected households per student per areal unit. There are tradeoffs among cost, identifiability, and sampling error that depend in part on the number of selections. (See Matsueda, Ross L. and William T. Bielby. 1986. "Statistical Power in Covariance Structure Models." *Sociological Methodology* 16: 120-158.) Likewise, there will be additional sets of data if the areal unit used to draw aggregate or random observations varies with the prevalence of demographic subpopulations in a locality.

Although starting with the matched ECLS-K, much new data will be required for the study. Since SES estimates and imputations will be conditional on choosing subsets of households that are similar to the student on various demographic characteristics, SES estimates will have to be re-calculated several times for each student (corresponding to different degrees of demographic specificity), for each level of geography.

The analysis will compare the validity of SES measures resulting from the 6 datasets by correlating them with ECLS-K parent survey measures and academic assessments and by comparing these correlations with those between ECLS-K parent survey measures and academic assessments.

[8] We may want to assign ZCTAs or combinations of ZCTAs to students where geocoding to lower levels of geography was unsuccessful.

Since the 2000 Census long form information is similar to that obtained from the ACS, this analysis will give us an idea of how best to create SES estimates for students in the 2007 NAEP SES pilot study, which is being conducted in the ACS Test counties.[9]

This ECLS-K analysis will also inform the SES pilot study by suggesting what the most feasible mix is of geographic levels and specificity of demographic matching and aggregation—taking into account rates of failure to geocode in detail and geographic variation in such failures. The analysis will also tell us what proportion of block groups, tracts and ZCTAs are too small (in terms of the number of matched households) to yield useful or reliable socioeconomic measures.

Based on these results, we will be able to refine and focus the methods of the 2007 and 2008 NAEP SES Pilot Studies. For example, it may be concluded that no SES estimates are possible at the block group level, since the ACS sample may be too sparse for this purpose.

Although the initial analysis will utilize the ECLS-K 3rd grade data (since this data set will most closely match the typical time lag associated with 5-year ACS estimates), other years of data could be analyzed, time permitting. Alternatively, one might argue that the lag in Census estimates is relatively unimportant and that analyses of ECLS-K 5th grade data are more relevant to the grade levels of NAEP assessments.

Several additional matters must be addressed in order to launch this study:

1. What data will we be able to recover from the Beveridge study? And how quickly? Can we get the student assessment data with matching block group, tract and zipcode (ZCTA) information as well as parent interview responses? We will need to obtain a license to analyze these data from NCES.

2. What is the best way to get new area SES aggregates and match household SES information, for each student at all three levels of geography?

    a. The Census Bureau creates the SES estimates according to our specifications.
    b. A WESTAT researcher works within the Census Bureau to create the SES estimates.
    c. Another researcher (e.g. Raghunathan at the University of Michigan or Hauser and Rivas at the University of Wisconsin) works at a secure Census data enclave.

    We think that it will be advantageous for project staff (UW or WESTAT) to create the measures, because one of the most important empirical questions is how to define appropriate demographic subpopulations within each areal unit in order to achieve sufficient sample sizes. For example, will we want to be as specific as, say,

[9] See R.M. Hauser and S. Rivas, "Multiple Imputation of Small-Area Data from the American Community Survey (ACS) to NAEP Student Records," rev. July 10, 2006.

households with a 9 year old white child, or—with the same analytic goal, say, households with a 7 to 11 year old white child? That is, we will probably need to vary the definition of population subgroups as well as the scope of areal units.

3. Will we have the time and resources to look at more than one ECLS-K year of assessment? If not, should the initial analyses be undertaken at the 3rd or the 5th grade levels?