Examining Small-scale Geographic Estimates from the American Community Survey 5-Year Data

Paper presented at the Annual Meeting of the
Population Association of America
Washington DC
March 31, 2011

Robert Kominski
Thom File
Social, Economic and Household Statistics Division (SEHSD)
U.S. Census Bureau

SEHSD Working Paper #2011-14

The authors wish to acknowledge the assistance and input of Census Bureau staff in the preparation of this research, notably, Michael Starsinic (DSSD)

INTRODUCTION & BACKGROUND

In 1996, the Census Bureau began the first field data collection associated with an ambitious new program, the American Community Survey (ACS). In December 2010, that program reached the realization of one of its major goals – the replacement of the once-a-decade decennial Census long-form data collection and delivery with data to be provided on an annual basis. This achievement marks an important milestone in the Census Bureau's long history, and has major implications for users of the data previously collected during the decennial Census activity.

The ACS program was not born of simple intellectual curiosity or the desire to try something new – in fact, as with many Census Bureau changes and improvements over the years, the ACS was stimulated and driven by very real fiscal, operational and data quality/data need issues. Some of the motivating issues are briefly discussed below – much fuller and complete discussion is presented in Chapter 2 of the ACS Design and Methodology report.[1]

After transitioning to sample household data collection in the 1940 Census (where some questions were for the first time not asked of all households), the Census Bureau evolved once again in 1960, when a mail-out form was introduced (instead of all-personal enumeration from the start of data collection). While costs were sizably reduced by implementing this mail approach, expenditures were curtailed further with the introduction of a 'short form' (with a small set of items required by the Constitutional reapportionment need sent to all households), and a 'long form' (the remainder of items necessary for a wide variety of statutory, programmatic and policy needs, sent to roughly one-sixth of all households).[2]

Thus, over about three decades (1930-60) the Decennial Census went from a fully enumerator-based activity with all households getting virtually all interview items, to a mail-out/mail-back format with only about 16% of household getting the full interview schedule.

Even in the era of the 1960's and 1970's, once-a-decade social, economic and housing characteristics data (i.e., those items collected in the long form) was often adequate to meet the various planning and policy uses for which decennial data were applied. During this time, few organizations outside of the Census Bureau actually had the hardware or software capability to work with the (then) massive datafiles produced by the decennial activity. A large part of the Census Bureau mission was then, as it had been for many decade, to produce information products - tabulations and reports – that summarized the decennial data results. By the late 1970's, however, advances in computing technology made it possible for more and more organizations to begin using decennial census data on their own, and for their own specialized needs. During this period the agency began to

[1] http://www.census.gov/acs/www/methodology/methodology_main

[2] For further details, see the Procedural History of the 1960 census at: http://www2.census.gov/prod2/decennial/documents/1960/proceduralHistory/1960proceduralhistory.zip

concentrate greater efforts on "public use" data files as another product for the general public.

One of the major utilities of the decennial census has been, and continues to be, its ability to deliver data not just for the nation as a whole, but also to provide detailed social, economic and housing information for very small-scale geographic units. This includes not only states and political subunits such as counties, cities and towns, but also units which to some degree are a hybrid of the operational collection units of the Census-taking activity, mixed with the notion of substantively meaningful small-scale geography such as neighborhoods and communities. In Census Bureau jargon, these small units are geographic areas such as Census tracts, block groups, and blocks. The point here is that over the decades beginning with the 1960's, these small-scale geographic units became increasingly important analytic building blocks for the growing army of data users with an expanding agenda of small-scale data needs and applications. More importantly, it was becoming increasingly clear that once-a-decade data and information, while generally reasonable in detailing the evolving nature of the national mass of several hundred million people, was inadequate in being able to provide monitoring information about rapidly changing areas and localities.

In summary, long-form census data, collected once every ten years, simply could not effectively address policy and research needs for rapidly changing areas as large as regions, states and cities. As major population shifts and social, economic and housing changes occurred throughout the back half of the 20th century, decennial census data, while an effective documentary of what had occurred, was less and less helpful for understanding trends already in progress, or those likely to occur in the future.

Thus, a variety of factors came together to cultivate a need for a different way of collecting and providing data and information that would fill the needs once met by the detailed (long form) Census. The American Community Survey represents the Census Bureau's vision of providing more timely and appropriate data to today's users, while also remaining adaptable to future research and analytical demands. As the ACS begins issuing its full array of planned data products, a high level evaluative question comes into focus. That is, can ACS data effectively replace (and even supersede) the data and information formerly provided by the once-a–decade decennial program.

This paper provides an examination of a practical applied problem (levels of high school dropouts) using publicly-accessible data and information from the first 5-year data file of the American Community Survey. The central question of this analysis is whether these data – intended to be the replacement for decennial-census long form data – are analytically useful in understanding this issue. That is, from an analytic perspective, can the small scale geographic information from the ACS provide useful information on this issue? Recognizing that the ACS is a continually evolving and improving program, we would nevertheless hope that even this first geographically detailed data set will be of value to researchers, analysts and policy makers.

DATA AND METHODS

The data for this project come from the first five-year file of the American Community Survey, covering the period 2005-2009. The basic design of the ACS is that each month, a quarter million surveys are mailed to households across the US and Puerto Rico. Follow-up for nonresponding mail households in a given month T occurs by telephone in month T+1, and then for continued nonresponding households, by personal interview for approximately one-third of this remaining sample in month T+2. The "one-year" datafiles for any given calendar year constitute the responding households for that year's data collection (a realized interview sample of about 2 million housing units), weighted to the July 1 population estimates for that year. Fuller details of this methodology are available in Chapter 4 of the ACS Design and Methodology Report. [3]

But the sample design of the survey supports "one-year" estimates that are reliable only for areas of about 65,000 persons or more. Thus, while these data products are useful for comparing states, large cities and virtually all metropolitan areas on an annual basis, they cannot provide reliable data at smaller geographic levels (i.e. any geographic unit with a population of less than 65,000 people).

To address this issue, the Census Bureau developed in its ACS design a series of other products which use larger amounts of sample data to provide reliable estimates for smaller pieces of geography. The first of these products uses data pooled over a 3-year period (referred to as the "3-year estimates") and was first released in 2008 (covering the period 2005-2007). Three-year estimates are then released each year thereafter, by dropping the first of the 3 data years and replacing it with the new 'current' year. This 3-year product allowed the Bureau to release data for geographic units down to about 20,000 people (or about 7,000 additional geographic units beyond the 7,000 initially available in the 1-year datafiles).

The final data product in this sequence is based on the accumulation of 5 years of data (that is, 60 consecutive monthly interview cycles). These data products ("5-year estimates") are intended to provide information for very small geographic areas, such as tracts and block groups, comparable in style to those that would have been obtained from the decennial census efforts of prior decades.

Preliminary research conducted by the Census Bureau has shown that the 5-year ACS estimates generally have larger variances than estimates derived from the decennial census long form, due in large part to smaller realized samples in the ACS. However, ACS data also generally have lower rates of unit non-response and item imputation, and the use of trained interview staff working on the data collection efforts continuously over time (instead of the approximately 6-month duration of long form data collection once every 10 years) ensures that the overall nonsampling error quality of the ACS data is better, and continuously being improved.

[3] http://www.census.gov/acs/www/methodology/methodology_main/

This research attempts to look at one substantive topic that might be the point of possible inquiry at a small scale level, using these 5-year ACS estimates that are available to the general public. We have purposely chosen a topic that is both substantively interesting and addressable using publically-accessible data products. Obviously, there are many topics of analytic interest for which the Census Bureau has not prepared external data products. One of our interests in this research is to demonstrate how challenging it may be for external researchers to try to answer a small-scale geographic problem with the data that are publicly available.

The issue of high school dropouts is one that has engaged researchers for some time. Since 1987, the Department of Education, by Congressional direction, has produced a yearly report on the level and characteristics of high school dropouts. While general levels of educational attainment have risen for numerous decades, high school completion, even for relatively young age groups, continues to generally not exceed 90%. For example, data from the 2009 ACS 1-year file shows high school completion for persons ages 25 and older to be 85.3% for the US overall, with the range across states varying from 79.9 to 91.8 %.

In some areas of the country, high school completion for young adults (all of whom might be thought to be well beyond high school completion age) can routinely be in the range of 75-80%. For example, in the 2009 ACS 1-year data, the high school completion rate for persons ages 18-24 in Baltimore city was 78.8%. Many public policy advocates and research foundations identify high school completion as one of several key national priorities.[4]

So, trying to understand dropout levels is clearly important, especially in the context of small geographic variation. Most children attend a school, which may be a 'neighborhood' school or part of their local school district. In some parts of the country a school district can be an entire city or county. In Maryland, for example, counties and Baltimore city, constitute all of the state's school districts. In other states a school district can be as small as a single school (this is common in places like Pennsylvania and Nebraska). Thus, making comparisons between high and low dropout rates at some point engages discussion of the schools and school districts children are a part of. Understanding the problem of high school dropouts – whether it be about schools or neighborhoods – is to some degree a local-level (small-scale geographic) problem.

Our approach in this analysis is to look at high school dropout levels for young adults in the District of Columbia. The choice is a conscious one – D.C. is a city of sweeping economic disparities and is characterized by both very wealthy and very poor neighborhoods. The city has undergone sizable demographic change in the past several decades, with an increase of White and Hispanic populations, and a decline of Black population. In 1970, the city was 71.1% Black, 27.9% White and 2.1% Hispanic (not a race); 2009 race estimates from ACS show it to be 53.2% Black (alone), 38.7 % White (alone) and 8.8% Hispanic.

[4] http://www.gatesfoundation.org/learning/Pages/2005-high-school-graduation-college-readiness-rates.aspx

Several years ago the mayor of Washington hired a new school chancellor to address the quality of D.C. schools, with one goal being the reduction of high school dropout rates. Trying to determine and understand dropout levels in a city like Washington, D.C. is exactly the sort of question or problem one might want to examine using small-scale data from the ACS.

The 600,000 plus population of Washington, D.C. is spread across 188 census tracts (a handful of these tracts have no population because they are completely within federal territory, and therefore have no residents). Figure 1 displays these tracts and their identification numbers. Census tracts are sampling and data collection areas defined primarily for census operation and survey sampling purposes. The Census Bureau does allow local area officials to help define tract boundaries, in an effort to approximate distinct defined neighborhoods or communities. The tract definition process is ultimately not a simple mechanical or statistical application. In many places (including D.C.), local officials have been involved over time in the boundary formation that readers will see in Figure 1.

The strategy in this analysis is to look at the published estimates of high school noncompletion in Washington D.C. tracts. In doing so we have not only chosen a somewhat rare event (or one we would prefer to rarely see), we have also chosen relatively small geographic units for examination. As such, this is a rather serious evaluation of the strength and utility of these small-scale ACS estimates.

We subject these estimates to both statistical and substantive criteria. The first evaluation – statistical – will be accomplished by examining the coefficients of variation associated with the tract level estimates. A coefficient of variation (or CV) is estimated as the ratio of the standard error of the estimate to its mean value. Over the past few years, the Census Bureau has internally discussed the utility of CV values to help guide users in assessing statistical quality of ACS data. Among the ideas previously discussed was a proposal to use colored-coded legends for all cells in ACS tabular data products, specifically in the American FactFinder data retrieval system, in an effort to easily convey to users a sense of the CV level for any estimate. While there is currently no 'official' Census Bureau (or other statistical organization) policy as to what constitutes a 'good' or 'bad' CV, the ranges that were discussed for a green/yellow/red "stoplight" presentation system would have used CV's of less than .15, for 'good' estimates, between .15 and .35 for 'moderate' estimates, and greater than .35 for 'bad' estimates (note that this is just one of many possible scales and researchers will debate their own opinions as to what constitutes a good or bad CV).

The second level of evaluation – substantive – cannot be as easily quantified. Just as there is no official document which definitely states that a given estimate is statistically 'bad', there is little direction which allows one to say a given estimate is or is not ***substantively*** good. However, beyond a quality assessment of 'good' and 'bad," there is a finer evaluative component of data being 'useful' or 'valuable'. It is not our attempt to create evaluated standards for these terms. However, one evaluative standard used commonly (including at the Census Bureau) is to attempt to analytically evaluate

'reasonableness' of estimates. One of the main ways this is accomplished is to look at a *collection* of estimates (not individual estimates) and to see how they compare to other collections of estimates. That will be the basis on the substantive evaluation, using other data from the ACS and prior decennial census.

Ultimately, from these admittedly rough standards, we hope to assess whether ACS 5-year estimates for high school dropout rates are 'good enough' for their intended purpose of evaluating educational attainment for small geographic areas.

ANALYSIS & RESULTS

Selecting the Data for Study

Since our approach in this analysis is to assume the perspective of a researcher outside the Census Bureau, the analysis begins by identifying a data item in the Bureau's existing array of products. One initial distinction of importance is the actual choice of data product type. The public use microdata sample files (PUMS) produced by the Census Bureau give researchers a great deal of control in identifying and constructing analytic variables and universes to specifically meet their research needs or interests. So, if for example, we wished to study the high school dropout status of young individuals who might have an English-language difficulty, and who also had some kind of physical disability, the relevant data items for identifying these persons are all available via PUMS.

The problem with this approach, however, is that the smallest available level of geography on the PUMS is the PUMA - Public Use Microdata Areas. These 'created' areas have approximate population sizes of 100,000, and generally do not delineate established political geography, such as a town, city or collection of such. Their primary purpose is to provide a relatively small piece of analytic geography for data users, without creating a serious data disclosure problem. Since our analysis of interest is for places smaller than 100,000 people, the PUMS is no longer a viable option.

This leaves the user with a large array of 'pretabulated' data, most of which is available in the 5-year dataset down to the level of Census tract (areas of approximately 2,000 households). In the 2005-2009 5-year file there are approximately 800 data tables that go down to the tract level. Since our analysis relies on small-scale geographic data, we need to identify a data table that comes closest to meeting our analytic need of assessing high school dropout rates.

A review of these tables provides several possible options. Detailed Table B14005 (Figure 2), provides information for persons 16-19 years old who are not enrolled in school and not high school graduates. This would seem to be the best option – it focuses on a relatively small age group in which dropping out is a critical problem, this giving us some idea of 'currency' in dropout activity. It is also disaggregated by gender, presumably allowing us to look at differences in this dimension. After examination, we

decided not to pursue this table, however, because sample counts were very small at the tract level, given the small size of the age group and disaggregation at the gender level.

Additionally, since the table expresses only estimated numbers and not percentages, effective statistical use of the data will require a recalculation of the margins of error into percentage terms, a process that produces an estimate of variance less accurate than the standard errors calculated directly from the original source data. Of course, the average public data user cannot do this. In future research we expect to return to this table, but for this exercise we have decided to try to find something with slightly larger data sample size, and with margins of error already calculated for the user.

Detailed Table B15001 (Figure 3) provides educational attainment information across 7 detailed categories for persons in various age categories (Figure 3). While this is a slightly larger age group, it still allows us to focus on young persons, because one of the age breaks is 18-24 years of age. Dropouts are actually enumerated across two categories of education, 'less than 9th grade' and '9th to 12th grade, no diploma'. (Note also that the table does not take current enrollment status into account, so in fact, some of these high school noncompleters could still be enrolled, although at ages 18-24, this is highly unlikely.) While the table allows fairly easy recombination of these cells (after downloading into a spreadsheet for data analysis program), computation of the combined standard error would again be possibly problematic for average public users. The Census Bureau uses a replicate weight formulation to compute the variances (and standard errors) of all published ACS estimates. These replicate weight variances are thought to be more accurate than those computed via standard simple random sampling formulae or generalized adjustments for 'design effects'. While it is possible to mathematically combine these published margins of error and recompute them as a single standard error value, this involves more downloading and recalculation (creating more chances for computational error by the user), direct computing (using the replicate weights) is a more accurate approach. However, this option is not available to the external data user.

We find a compromise solution in a data element that is available though one of the 'Subject Tables", also available in American FactFinder. Subject Table S1501, Educational Attainment (Figure 4) provides in one place a collection of data items taken from the Detailed Tables. One of the cells in this table provides the high school noncompletion percentage (i.e. – dropouts) for persons ages 18-24, with a directly computed standard error. In the context of the kind of person we are simulating – an analyst in a local education or funding office who has been asked to provide data perhaps in an hour or two – this is would probably be their best and most efficient choice (the researcher could get a single estimate, with an already calculated standard error, via a user friendly application like American FactFinder). It is with these thoughts in mind that we settled on Subject Table S1501 as the data source for our analysis.

Step 1 – ACS HS Noncompleters Ages 18-24

Map 1 shows the percentage of persons, ages 18-24 years old, who have not completed high school for the tracts of the District of Columbia. These values, which were taken directly from Subject Table S1501, range from a low of 0% to a high of 81.9%. These estimates, along with their corresponding number of sample cases and computed coefficient of variation are detailed in panel 1 of Table 1.

While the map is characterized by large amounts of blue (low levels of high school dropouts), there are some tracts in yellow to red color, indicating point estimate dropout levels of 50 percent or more. Many tracts in Northwest D.C., an area of relatively high incomes, low poverty and high adult educational attainment, are blue. Areas in the lower right hand corner, including areas which have high poverty rates, high unemployment and generally lower overall educational attainment, are progressively less blue. Note that a handful of tracts (mostly federal property, colored black) have NO data and cannot generate an estimate at all.

Of course, these point estimates are driven by the collected samples, both in terms of size and representativeness. Map 2 shows the number of sample cases that underlie each estimate. As can be seen, the vast majority of tracts have 20 or fewer sample cases (in their denominator) and 41 of them have an estimate of zero, which might truly be their estimate, or might just as likely be a function of insufficient sample. Regardless of the zero values, the estimates in general are being driven by very small samples – this is demonstrated more clearly in Map 3, which shows the estimated coefficient of variation (CV) for each tract estimate. With no strict guideline of a 'good' or 'bad' CV, we have arbitrarily used breaks of .5. A sizable number of tracts (41) have no CV because they have a zero estimate; a handful (13) have CV's above 1, and 52 (of the 175 tracts with data) have an estimated CV below .5. Map 3 clearly draws into question the statistical quality and utility of these data. What is not yet clear, however, is how these estimates look in the context of any other data that are available, or that one might use as a referent standard for evaluating them.

Step 2 – ACS High School Graduates, Ages 25+

The previous analysis indicated fairly clearly that while the ACS data can and do provide a high school dropout estimate for a relatively small age group for tract data, those estimates need to be taken with great caution, statistically speaking. Of course, one might improve the estimates if we simply had larger samples upon which to base them. While we cannot go back and get more sample, we can once again adjust the universe of interest to find a measure that includes more sample and thus, hopefully, provides a better indication of the level of differentiation in high school dropouts across the tracts in D.C.

Luckily, such an estimate also exists in Subject Table S1501 (Figure 4), where the percentage of *all persons ages 25 and above* who have completed high school, is provided. Obviously, from a substantive standpoint, this estimate is very different than

the one just discussed. In addition to the sizable change in age universe, we have now reversed the concept itself, changing the focus from high school dropouts to high school graduates. This is because the estimate in the table is about high school graduates – not dropouts. And again, while the more analytically appropriate estimate could itself be simply obtained (1-graduates = dropouts), and the standard error is actually the same for both estimates (since se(p) = se(1-p)), how likely is it that a planning commissioner, or their assistant, asked to have data for a meeting or presentation in a hour or a day, will have the knowledge or motivation to do this? The easier and more likely path is to simply use the number that is clearly available.

Map 4 provides estimates of the percent of persons ages 25+ who are high school graduates (or more) for the D.C. tracts using the same 2005-2009 ACS data (actual estimates in panel 2 of Table 1). As can be seen, the picture changes somewhat. Noting that the color gradient has been reversed to correspond conceptually to the first set of maps, much of the map is now either dark or light blue (representing estimates in the range of 80+% high school completion). Only one tract – 60.02 – is colored yellow on the map – the estimate of 35.3% high school completion is based on just 18 sample cases, and has a CV of .65. Maps 5 and 6 show, respectively, the number of sample cases and CV's associated with these estimates. While in the previous analysis the majority of the 188 tracts had 20 or fewer cases, in this map there are now just 14 tracts with fewer than 20. Just 11 of the 188 tracts have fewer than 50 sample cases. Correspondingly, the CV values also drop dramatically. Only 2 tracts have a CV of .25 or *greater*; in the previous analysis just 12 tracts had a CV *below* .25. The vast majority of the tracts now have very small CV's – 170 tracts have a CV of less than .10.

The switch to this estimate is clearly a statistical improvement over the initial measure. However, from a gross evaluative standpoint, large sections of the map are in concurrence with Map 1. Northwest D.C. is solid blue (good) in Map 4, and nearly so in Map 1 as well. Color variation, indicating variability in the estimates, is most evident in the tracts running along the bottom right border (Southern Avenue). In short, as much as of an improvement that Map 4 is from a statistical standpoint, it reinforces many of the basic substantive conclusions one might reach by looking at Map 1 alone.

Step 3 – 2000 Census High School Graduates, Ages 25+

One of the beliefs about the routine production of ACS data is that it gives users the new opportunity to begin tracking phenomena, both across geography and time, on more than a once-a-decade basis. As ACS estimates have begun to be produced, it has become common practice for users (including internal to the Bureau) to provide comparisons of ACS data estimates with those taken in the last long form census of 2000. These comparisons began as early as 2000 itself, when the Census 2000 Supplemental Survey (C2SS), the first national expansion of the ACS, was used to provide estimates and evaluation reports comparing these two data sources.[5]

[5] See, for example, http://www.census.gov/acs/www/Downloads/library/2004/Report09.pdf

While the decennial census and the ACS have many differences in design, format, implementation and even question content, in many cases, estimates between the two surveys are felt to be highly or moderately comparable.[6]

In this section, we look at the estimates of high school completion measured in Census 2000. As with the previous analysis, the measure of interest is the percentage of persons ages 25 or older who have completed high school. These data are also easily obtainable via the American FactFinder system, so it is not unreasonable to expect that someone studying the dropout problem might wonder how, if at all, things have changed in D.C. in the last decade.

Map 7 provides the estimates for high school completion from the Census 2000 for the D.C. tracts (Table P37). Visual comparison of this map to Map 4, using the 2005-2009 ACS data, shows a high degree of visual similarity. Of course, the simple tract-to-tract visual comparison also shows that there are numerous points of difference between the estimates across the two maps. The 'west' side of the map is nearly as blue as it is for the ACS data. Moving from left to right, there is also a fair amount of variability through the center of the city, just as we saw in the 2005-9 ACS map. Of course, over the past decade, Washington D.C. has seen lots of changes as the city has become more prosperous and ethnically diverse. So, it isn't unreasonable that the two maps do not look exactly the same.

But from a substantive standpoint, it is not amazingly different. A statistical test of the Census 2000 estimates compared to the 2005-2009 ACS estimates shows that just over half of the estimates, (99 of 182 comparable tracts) are not statistically different at the 95% or greater level.

Table 8 shows the total number of sample cases (not just age 25+) for tracts in Census 2000. As can be seen, while most tracts have well over 80 sample (long form) cases, a handful do not. Using the documented generalized parameters for Census 2000, we were also able to compute CV's for the tract estimates, shown in Map 9. As can be seen, the CV's are remarkably similar to those obtained from the 2005-2009 ACS data (Map 6) for a comparable concept and universe. While 170 of the tracts in Map 6 had a very low CV (below.10), in Map 9, 180 – nearly all – of the tracts have a CV below .10. This comparison is perhaps the most encouraging of this entire analysis. A major criticism of the ACS is that the data that will come from the survey for small scale geography will not be as "good" as the data that came from decennial censuses past. The real unexamined issue may ultimately be not how 'good' ACS 5-year data are, but how 'bad' long form tract data often were.

Step 4 – ACS High School Graduates, Ages 18-24

Early on, in order to get more sample and a better base of comparability to existing Census 2000 results, we chose to switch to the full adult population and the more common phenomenon (high school completion versus dropping out). But given the

[6] See, for example, http://www.census.gov/acs/www/guidance_for_data_users/comparing_data/.

results we have achieved, one can wonder what happens if we keep the more common measure (high school completion), but return to the population of greater relevance (young adults). ACS Subject Table S1501 has just such a measure – the percentage of persons, ages 18-24 years old, with at least a high school diploma.

Map 10 displays these estimates. As with Maps 4 and 7, the color patterns, while not matching exactly, are similar. While the number of relevant sample cases remains the same (Map 2), Map 11, showing the estimated CVs, is far more encouraging than Map 3, and very similar to Maps 6 and 9. This set of estimates, based on 5-year ACS data for a relatively small subpopulation (18-24 year olds) at tract level geography, shows interesting substantive variation across the city, *and* has a high degree of statistical reliability. Just 6 tracts have a CV of .5 or greater, and 58 have a CV of less than .10. In short, examining virtually the same concept at the tract level, even for the small subpopulation of the age group of 18-24 year olds, yields results which are not strikingly different from those in Maps 4 and 7.

From both a statistical and substantive standpoint, Maps 10 and 11 represent quality small-scale geographic data for an important substantive topic and subpopulation. We believe that these results could be reproduced for a wide variety of other social, economic and housing phenomena.

As a final 'reality check', Maps 12 and 13 show the side-by-side comparison of the estimates and CV's associated with our original measure (persons 18-24 with less than a high school diploma) and the final measure (persons 18-24 who have a high school degree). In Map 12 the substantive similarities across the city in terms of estimates are quiet apparent. In Map 13, the sizable difference in statistical quality is equally striking.

One Final Check

Quite often, social scientists use a given empirical measure not because it is substantively 'the best' or most appropriate, but simply because it is the measure most readily available. In some cases, these inferential leaps may be more than appropriate. One simple test of the utility of alternate measures is to try to assess how well the alternatives correlate with the measure we would prefer to use. The table below shows the intercorrelation matrix of the four different measures used in this study.

The items are:
Measure 1 -- 2005-9 ACS Dropout level, ages 18-24
Measure 2 -- 2005-9 ACS High school completion, ages 25+
Measure 3 -- Census 2000 High school completion, ages 25+
Measure 4 – 2005-9 ACS High school completion, ages 18-24

	M1	M2	M3	M4
M1	*	-.520	-.525	-1.00
M2		*	.826	.520
M3			*	.525
M4				*

(Note: This analysis was run both including and excluding the "zero estimate" cases. Especially with regard to Measure 1 (dropouts) an estimate of zero may simply be a result of not having enough sample. In a case such as that, the zero estimate is misleading. Results from the two analyses vary only slightly – the less 'robust' estimates (with zero cases omitted) are shown here).

The original measure, M1, correlates moderately well with both the expanded age measure from ACS, and the Census 2000 measure. In turn, M2 and M3 (ACS and Census2000) also correlate fairly strongly (.826), reinforcing the notion that even with changes over nearly a decade, the ACS data corroborate the patterns seen in the 'better' Census 2000 data. And, since M4 is simply the converse of M1, these correlate perfectly at -1.0. In short, all four measures provide fairly strong associations, both across time, datasets, and age ranges used.

SUMMARY & IMPLICATIONS

This research has had multiple purposes. One objective has been to illuminate the utility of small-scale geographic estimates from the 5-year ACS data. As the intended replacement of the decennial census long form data collection program, the ACS has to show that it can be as useful as its predecessor. In some cases, it may actually have to show that it is even better. Some applications, such as using the data over repeated cycles (using overlapping samples) may well eventually help to demonstrate modeling and synthetic data applications that many users have not even yet fathomed. Time – and innovative data users – will help create many unimagined uses of ACS data in the years ahead, just as happened with decennial census and other data collections that continue to have applications far beyond what anyone initially might have thought, or designed.

Another purpose has been to show that users will have a large number of choices and limitations that they will need to deal with in finding appropriate ACS data to answer their substantive questions and problems. This includes finding ways to use the data that are reasonable and responsible. Not every user will be thorough enough, or statistically or computationally adept enough, in finding ways to best use the data to answer the questions they have. Hopefully, over time, a body of research and application examples

will evolve to help establish best practices and important methods which will help to guide users more efficiently to good analytic decisions.

A final and perhaps main objective has been to demonstrate, through the use of a single case example, the statistical and substantive quality of a typical practical example that ACS data are likely to be used for. The example used in this research demonstrates that while users need to be aware and mindful of the statistical properties of these small-scale data, they nonetheless have useful substantive value – probably as much as decennial census long form data did, and in the long run, when there are 10 or 20 years of annual sequential estimates available, perhaps far, far more.

Will users have to examine ACS data products and metadata to get a good understanding of the variety of choices they have, and the value and limitation of each? Yes.

Should analysts show caution in looking at specific tract estimates and making comparisons? Yes – just as they should have when using long form census data.

Are small-scale geographic (and subgroup) ACS data of substantive analytic value? More research is needed, but the results here are very encouraging.