October 2, 2010

To: Interested researchers using National Student Clearinghouse data

From: Sara Goldrick-Rab and Douglas N. Harris, University of Wisconsin-Madison[1]

Re: Observations on the use of NSC data for research purposes

---

The purpose of this memo is to convey some observations and lessons learned from the use of National Student Clearinghouse data in several projects, most notably our ongoing randomized trial of need-based financial aid, the Wisconsin Scholars Longitudinal Study (www.finaidstudy.org). In the WSLS, Clearinghouse data are used to measure the key outcomes of college persistence, transfer, and degree completion. We have worked with Clearinghouse data for a cohort of 3,000 Wisconsin students attending 2-year and 4-year colleges across the state, beginning in fall 2008.

The NSC is a non-profit organization founded in 1993 that serves as the nation's only source for college enrollment and degree verification. It is a centralized reporting system which collects publicly available directory information obtained from the colleges and universities attended by 92 percent of American undergraduates. Over time, more institutions of higher education have joined the Clearinghouse, reducing an important concern that enrollment would be missed if it occurred at non-participating institutions.

Recent years have seen major investments in the NSC, including a nearly $3 million grant from the Bill and Melinda Gates Foundation intended to expand its StudentTracker system. Absent a national student unit record data system, the NSC is the only game in town for researchers wishing to know where students in their studies attended college—particularly if those students crossed state lines or attended private colleges. The NSC is playing an important role for school districts and institutional researchers at colleges and universities as well, since it enables them to estimate the college participation, transfer, and completion rates of their students.

But the effective use of the NSC, as for all sources of administrative data, depends heavily on the researcher's ability to make sense of how the data is assembled and properly used. The NSC is an advanced system, but it is not as clean or complete as it might seem. Moreover, most "talk" about the NSC among researchers relates only to the issue noted above—which colleges participate.

[1] We are grateful to NSC research director Douglas Shapiro and his staff for conversations about their database and its use. We also thank WSLS staffers James Benson, Robert Kelchen, Lisa Klein, and Peter Kinsley, for their substantial assistance in our assessment of the NSC data.

There are other important aspects of the NSC researchers need to attend to when using its data, and that is what we hope to lay out here.

The first issue regards how the NSC records college enrollment. After checking to be sure that the data file received from the NSC contains a list of all students named in the requesting file, the researcher should next examine the "Record Y/N" indicator. This indicator is the primary measure of whether college enrollment took place. If Record=Y, the student was enrolled in college for the period of time indicated in the next columns, at the institution indicated. If Record=N, this is meant to indicate that the student was not enrolled in college for the time period requested.

The problem is that the latter assumption rests heavily on NSC's capacity to query all possible records for college enrollment of said student. While we know that enrollment at non-participating institutions could be missed, or that a student's record may be absent from the database because it is blocked, what researchers must attend to is the possibility that Record=N means the NSC was unable to match the student to her record contained in the database. [2] Inaccuracies in matching could understate overall enrollment rates and introduce measurement error.

Thus it is imperative that researchers better understand how the NSC matches students to records. The process usually requires submitting a file of students' names and dates of birth to NSC, which then executes a matching algorithm that sometimes invokes probabilistic, or "fuzzy logic" to allow for possible errors, such as typos, either in the requesting data or in the data that NSC has received from the colleges. In some cases (where allowed under FERPA), the NSC uses social security numbers; however, we have evidence from at least one research team that the SSN search has resulted in some inaccurate matches (e.g. if 3 or 4 digits match, then record=Y; a likely false positive) whereas the opposite seems to occur for name/date of birth searches (e.g. in our experience, if even one letter is different in a name, then record=N; a likely false negative). In our sample three percent of students initially received false negatives due to very small differences in their identifying information as contained in our request versus what was in the NSC records. The problem with this reliance on accurate names and dates of birth is that administrative data and self-report data on such identifiers are often flawed—students regularly use different information on

[2] By law, students may opt to block their enrollment information, resulting in the absence of their records in the NSC database. Colleges are also allowed to block records. The control report returned by the NSC to the researcher indicates the prevalence of blocked records.

their FAFSA and college record and in surveys. Given that this is a national census-like database, students can also share the same names and dates of birth.

The matching process also varies according to the type of search the NSC processes. Researchers have five different search types to choose from, but most often use two options: a cohort query and a subsequent enrollment query. While information on the specific times and situations for each is explained on the NSC website (http://www.studentclearinghouse.org), results can vary depending on which type is executed. The search algorithm utilizes the requesting college as a starting point in a cohort query, which can result in false negatives when the search doesn't find the student in that college. For example, in our study a subsequent enrollment query was able to detect enrollment for nearly six percent of the sample that went undetected by a cohort query.

For these reasons, researchers might be tempted to use the NSC in conjunction with another administrative data source, such as a university system's data. In this case, they would be advised to consider how enrollment terms are defined by the NSC. Specifically, while most colleges and universities measure enrollment beginning at the start of the term, the NSC receives the first enrollment records from colleges anytime up to 30 days after the start of term. Some colleges submit earlier than others and some have terms of differing lengths. Thus, a student who began college September 1 but withdrew by September 30 would not have an NSC record for that term if the college had not yet sent its data to the NSC at the time of withdrawal. Clearly this could have a disproportionate impact on studies of students with more tenuous holds on higher education.

We also note that while we have not yet analyzed these issues in our own research, we have been told that while the NSC's data on enrollment intensity and degrees completion are increasingly available, these data may vary in quality not only by college but also by students within colleges.

So, what can researchers do to become more confident that they have identified correctly whether students in their study are enrolled in college? In addition to communicating with the very responsive NSC staff, we recommend the following steps:

(1) When constructing your request file, obtain identifying information on students from multiple sources. Make sure that at least one of those sources is the student's college, if possible (the NSC data is based on the college's records). Submit in the request to the NSC all variations on names you have for a student. That is, if you have three different names, then include three rows for that

student in the requesting file, one for each variation. Do the same for variations in the dates of birth and SSN (if you are allowed to use SSN).

(2) Because the NSC usually charges researchers for requests, few can afford to submit multiple requests to assess data consistency and quality. But we caution that when possible it is advisable to budget for the financial resources and time necessary to submit repeated requests to the NSC for the same students and the same enrollment periods to make sure the search algorithm is working as well as possible. For example, in January 2010 submit a request for 3,000 students, querying enrollment in 2008-2010, and do the same thing again in June 2010. Compare the results. Ideally, try and submit the query using the college or system at which the students in your study originated. When you receive a file for a more recent term, re-examine the data in that file from older terms—do not simply append the new term. But keep in mind that the NSC data can change over time – colleges submit several enrollment files per term so records can be added or removed, and students can add or remove FERPA blocks. Also, the NSC works to enhance its matching process over time, so matching results, particularly for imperfect data, can vary.

(3) If you know for certain that the students in your study were enrolled in college, then indicate in your paper the rate at which the NSC found a matching enrollment record (e.g. the "hit rate"). This will help our community of researchers a sense of common hit rates for different populations. In our study, we knew from state administrative data that all of our students were enrolled in college as of the start of September 2008, but after repeated requests the NSC produced college enrollment in that term for 96.3 percent of the 3,000 students (this included 97.5% of 4-year students and 95.2% of 2-year students).

(4) In all types of studies using NSC data, consider that the descriptive statistics about college enrollment, persistence, and graduation still likely under-state the true levels. In addition, note the following issues when estimating causal impacts of programs and practices using NSC data. When the NSC's enrollment reports are used to create dependent variables for causal inference, it will often—but not always—be reasonable to assume that the measurement error is orthogonal to treatment condition. Specifically, if selection to treatment is non-random, the selection process might be related to the quality of identifying information. For example, more disadvantaged students may be more likely to fill in identifying information incorrectly on the FAFSA and other forms, resulting in a lower NSC match rate for these students. If disadvantaged students in this scenario are also more (or less) likely to select into the treatment, then the measurement error is no longer orthogonal. If you have an alternative data source for college enrollment, you can and

should look for orthogonality by testing differences in hit rates (if an additional data source is available; see above). Otherwise, orthogonality should be stated as an assumption.

In random assignment studies, the orthogonality assumption will be highly plausible, but even in this situation it depends on the timing of the treatment. If the treatment comes before students could be expected to attend college, then the treatment could affect the quality of information available for the NSC request. For example, a program that helps students fill out the FAFSA might make them more likely to attend college *and* more likely fill out the form correctly. If the FAFSA data are used to create the match in this situation, then this will increase the hit rate in the NSC data and bias upwards any impact on enrollment. A higher proportion of the control group will have enrolled in college, but not been found in the NSC. Therefore, in studies of pre-college programs researchers need to demonstrate evidence supporting the additional assumption that the treatment does not affect the NSC requesting information.

Regardless of randomization, the situation is different when the treatment examined occurs during college. If the treatment begins on or after the second semester of college, one can test the orthogonality assumption by comparing enrollment rates in pre-treatment semesters. They should be equivalent in a randomized trial. In a quasi-experiment, this prior enrollment information might be unequal, and prior enrollment would become an independent variable (albeit one with some measurement error and attenuation bias).

The situation is most complicated in studies like ours where the treatment begins during the first semester in college. In that case, whether we have a valid test of orthogonality depends on exactly when the treatment started vis-a-vis date of record. In these cases, the dates of enrollment contained in the NSC report make it possible to know not only in which semesters students were enrolled but also their specific dates of enrollment. Whether this additional information is useful depends on whether the researcher also has data on the precise timing of the treatment.

To summarize, our intention in this memo is not to discourage the use of the NSC—in fact, as we suggest earlier, it is a terrific and uncommon resource. At the same time we feel it is important to draw the attention of the research community to the vagaries of the NSC as a data source for studies where precision in outcomes is important. As with all data sources it contains flaws to which researchers must attend. Nevertheless, it represents one of the best available measures of student enrollment across institutions of higher education.