Conversation between Professor Michael Rosenfeld (Stanford) and Holden Karnofsky (GiveWell) - August 21, 2012

Summary:

We spoke with Professor Rosenfeld about meta-research because we were impressed that he made the data from his "How Couples Meet and Stay Together" study public. We also spoke to him about his research on relationships because we perceive this to be a subject on which relatively little research has been done, and one where further research could potentially have a significant social impact. Some key takeaways:

- Prof. Rosenfeld laid out a variety of obstacles to data sharing, including the challenges of confidentiality and the challenges of making data usable by the general public, and also discussed some of the efforts underway to make data sharing more practical and more common.
- Prof. Rosenfeld also discussed his research and how it fits into the bigger picture. He discussed limitations of self-reported/survey data, and what he believes is known (and what he is interested in researching more) about the factors behind couples' forming and staying together.

Full notes:

This is a set of summary notes compiled by GiveWell in order to give an overview of the major points made by Michael Rosenfeld in conversation.

Meta-Research:

Challenges of data sharing

It's often the case that researchers lose their data and so are unable to share it with other researchers. It's also often the case that researchers don't clearly explain how they performed calculations with their data to arrive at the results in their papers. One of the issues is that researchers feel vulnerable to the possibility that if they share their data, somebody will find flaws in their work, and so are reluctant to share their data. Other issues pertain to confidentiality.

When social scientists gather survey data, they generally promise the participants that their results will be kept confidential. In the modern era, it's becoming easier and easier to identify individuals based on their survey responses even if the obvious identifying traits such as name and address are stripped from them. This is because it's increasingly

the case that people have access to large databases containing many identifying traits of individuals. This creates an obstacle for researchers who are interested in making their data sets public. At times, I've had to be careful to omit certain information such as the state that the subject lives in in order to avoid tacitly compromising confidentiality.

My "How Couples Meet and Stay Together" (HCMST) project collected 3000 stories from people about how they met their partner. This sort of collection of stories is a useful complement to samples of interviews and large multiple-choice surveys. However, in order to share it with other researchers, my team had to go through each story and edit out all features that might identify the writer. This took us several months. We did not make the collection of stories public: researchers need institutional review board approval in order to access the text answers.

There is a set of federal laws about the confidentiality of medical data, and these laws can be quite restrictive.

There are some working papers at the Interuniversity Consortium for Political and Social Research (ICPSR) website giving guidelines for data sharing. The Census Bureau also puts out literature about the maximum number of variables that should be shared from a data set of a particular sample size. Aside from these, there aren't standards for how to prepare data for public sharing. I think that one reason for this is that the variables that others need to identify people are constantly changing.

Researchers often put together their data sets for their own usage rather than for the general public. Reformatting data sets to be understandable to others can involve a substantial amount of work. This is especially the case if the researchers do so a long time after they originally compiled their data sets, when the meaning of the data is no longer fresh in their minds. This work can be partially avoided if researchers put together their data sets with a view toward eventually sharing them with the public from the outset. This was the case for the data set from my HCMST study.

Data sharing and software

The most popular data compilation software in sociology, economics and perhaps epidemiology is STATA. The license for STATA costs hundreds of dollars and the program is difficult to use. These facts pose an additional challenge for researchers interested in sharing their data with the public, and for members of the public who want access to researchers' data sets.

There is a useful website maintained by UC Berkeley called Survey Documentation and Analysis (SDA) which offers data analysis tools that are easy to use and that can be used

for free. The website might a good resource to fund if it has funding needs.

When I did my HCMST project, I solicited help from the Stanford University Libraries to help me build a website to communicate my data in a way that

(i) Users could interact with easily
(ii) Makes it easy to contact users about updates
(iii) Makes it easy to keep track of how many people were using it.

This assistance was very helpful. I was able to get my data out to the public fairly quickly.

Helpful organizations

Much of the data that researchers compiled before the digital age is on microfilm and so is inaccessible. There is an organization called Integrated Public Use Microdata Series (IPUMS) that was created by Steven Ruggles at University of Minnesota. IPUMS converts old microfilm data from (i) US Censuses, (ii) international censuses and (iii) other large-scale data collection efforts into digital form, and makes this data accessible. IPUMS is a very helpful service.

IPUMS received around $50 million from the National Science Foundation (NSF) and from the National Institutes of Health (NIH) for this project. That said, if the organization still has funding needs, it could be a good candidate.

Interuniversity Consortium for Political and Social Research (ICPSR) offers a great service, but has some limitations:

1. It takes them a long time to process data that researchers send them and put it online. They took substantially more than a year to process my HCMST data, and for a while, whenever I updated the data set, it took them a year to update their site with it.

2. ICPSR is not free to members of the public. I understand that it costs a lot of money to have staff do all of the vetting that they do, and that they have to charge somebody for it. But I wish that their product were free to the public.

Ideas for mechanisms to promoting data sharing

As you wrote on the GiveWell blog, one way to promote data sharing is to get journals to require that authors upload the data sets and/or the code that they used to analyze the

data. Jeremy Freese is a sociologist at Northwestern University who is working toward influencing journals to do this. There are relatively new journals in sociology and in economics that are tackling questions such as how to get prestigious journals to focus more on data and less on technical analysis.

In practice, it's hard for journals to determine whether the data and code that the authors provide is enough to replicate the analysis without replicating the analysis in full themselves. Of course, doing this is very time consuming.

My impression is that the NSF is trying to be more rigorous about making sure that grantees follow through on their promise to share their data. They have a new requirement about data management plans: when a researcher applies for a grant from the NSF, he or she is required to submit a data management plan connected with the proposed research project.

Funders could hold grantees accountable for sharing their data by renewing only the grants of those grantees who share their data.

It might help to give recognition to the institutions (e.g. IPUMS, ICPSR & SDA) that promote open data.

Motivations for researchers to share their data

The more people use a researcher's data, the more important the researcher's project will be. This can provide an incentive for researchers to share their data. I think that researchers should give greater consideration to the possibility that sharing their data will increase their reputation.

Some researchers want to contribute to learning and share their data or course materials for that reason.

Putting out more material opens one up to more criticism, but I find the benefits of making my materials public to be well worth this cost.

Publication bias

Publication bias is a large problem in the social sciences.

It's frequently the case that many people believe a view because it's become accepted. This makes it easier to publish findings that agree with the accepted view. This results in the accepted view being reinforced even when there is no additional evidence for it being

true.

If a paper is in a highly specialized area, often the paper's reviewers will be the researchers who originated the accepted view, or the students of the view's originators, so that the reviewers are motivated to be skeptical of findings that contradict the accepted view and reject such papers disproportionately.

There is also publication bias in the direction of publishing studies that *employ fashionable tools* being published more frequently than those studies that do not. This provides incentive to researchers to use fashionable tools even if they're not the best tools to use to research a given subject.

Other comments on funding

The NSF has been very generous with me. I really appreciate that they were willing to fund me to do my own project on its own merits even though I was young and didn't have very much experience in the area.

Some funders other than the NSF want to know the results before funding a project. This incentivizes researchers to work on projects that have already been done.

Research on relationships

The inadequacy of survey data

Survey researchers have found that answering questions on surveys well requires more cognitive ability, commitment and energy than most subjects have to give. People frequently don't understand survey questions and so will pick answers randomly rather than spending time trying to understand the questions.

Nontraditional families

I'm especially interested in the study of interracial couples, interreligious couples and same-sex unions. I wrote a book titled *The Age of Independence: Interracial Unions, Same-Sex Unions, and the Changing American Family* about this subject.

Nontraditional couples tend to live in a state different from where they were born whereas traditional couples tend to live near where they grow up. Sometimes people's communities don't accept their relationships when they marry someone of a different race or religion, or when they enter into a union with somebody of the same sex. This prompts them to move to a different location such a big city, where the community is more

accepting of their relationship.

I believe that it's important to get good data about gays, lesbians and bisexuals in order to inform policy.

My research on same-sex unions attracted controversy when a Republican senator from Oklahoma highlighted the grant that I received from the NSF as an example of wasteful government spending on research. My research gets negative attention from people whose view on gay rights differs from my own.

Relationship formation

My book titled *The Age of Independence* gives a thorough survey of the literature on relationship formation in general.

In 2005 I wrote a paper presenting evidence against a theory called status-caste exchange. This theory is that, for example, a white person would only marry a black person if he or she were "superior" to the white person in some respect (such as wealth, social status or education). The theory originated in the 1940s and was not supported by evidence at the time. In the 1970s researchers looked at data and found that they were was no evidence for this theory and so the theory was discredited. In the 1980s and 1990s researchers used sophisticated statistical methods to analyze the data and arrived at the conclusion that the evidence supports status-caste exchange. In 2005 I found that their analysis was largely erroneous, that the analysis that the researchers from the 1970s did was sound, and that there isn't any evidence for status-caste exchange.

This is an example of a situation in which researchers were led astray by being attached to fashionable complicated models. A lot of the researchers in my field still have this attachment and so believe in status-caste exchange.

Divorce and break-ups

There is a substantial literature on divorce in sociology. This is partially due to the dramatic increase in divorce rates in the United States during the 1970s.

People who get divorced are unrepresentative of the general population and so it's difficult to infer a causal relationship between observed variables and divorce. However, there are some suggestive correlations.

- Early data suggested that children who had gone through divorce tend to be

disadvantaged relative to other children. This raised the possibility that parents' divorce causes their children to be disadvantaged. However, new data suggests that the children whose parents divorced were not only disadvantaged after the divorce, but also before the divorce, suggesting that the disadvantage is not caused by the divorce but rather by other distinctive characteristics of members of a family where the parents are prone to getting divorce. The question of what the effect of divorce is on children is an unsettled question.

- According to the literature on divorce, people who are poor, who marry young, and who marry someone of a different race or religion are more likely to break up than other couples. The data that I've recently collected does not support these hypotheses. While interracial are interreligious relationships are less likely to form than one would expect by chance, those that do form appear to be just as likely to last as other people's relationships.

- I have a recent paper finding two predictors of relationships lasting:

    1. Couples that have been together for longer are more likely to stay together in the future

    2. Couples that have entered into a public formal union are more likely to stay together in the future (regardless of whether the union is sanctioned by any particular authority).

    The frequency of break-ups among people who have been together for a few years and who have had a formal union is 1.5% per year, which is relatively low.

In the future, I'm going to be doing a lot of work on the predictors of break ups. One thing that I'm interested in is studying the relationship between self-assessment of relationship quality and chances of breaking up. We've just started a longitudinal study and asked the subjects how they rate the quality of their relationship.

We also asked them what they attribute the quality of their relationship to. We've learned a lot of interesting things that we wouldn't have thought of from what they wrote. For example, something we've learned from this is that many people characterize their relationship as great because both they and their partners have relationships with God.

I tend to be skeptical of self-reported measures of personal psychological profiles because in my (limited) experience, people seem unable to offer accurate descriptions of their personality traits. So I haven't attempted to incorporate such metrics into my studies.