Jon Krosnick: Thank you Bob. Well not knowing exactly what Bob was going to talk about it turns out that this sequence here to our next presentation is perfectly suited. Josh Pasek at the University of Michigan has done I think what Bob is suggesting we should do in a small way linking survey data to auxiliary data and he will tell us about the experience that he has had there and the context of the larger literature of these efforts so far.

Josh Pasek: So unsurprisingly given the extent to which we’ve been sort of talking about this so far, there hasn’t been a ton done trying to assess exactly what we end up with when we merge together these big datasets and survey data and start looking at quality. And so what I’m going to show you here is sort of an initial data cut at accuracy and bias in the relations between consumer file marketing data and what we end up getting off of in this case GFK knowledge panel as they’re recruiting individuals into their sample. So first to reiterate Bob a little bit, I sort of put up a couple of the challenges and opportunities I think we’re dealing with as a field ‘cause they’re a good context in which to start.

And our challenges as we’ve all sort of discussed and thought about are declining response rates. We have the Pew numbers there. We’ve seen a number of other things that sort of note how response rates have been going down over time. We’ve seen increasing costs over a variety of different modes for various reasons, increasing refusal, increasing difficulty to reach particular populations et cetera. And for some modes at least we’re getting more and more coverage challenges which it’s becoming more and more difficult to reach Hispanic and young Americans in particular and how do you get those kinds of individuals into the sample. Issues like who has telephone access and how we deal with cell phones have raised costs, have introduced additional challenges into what was for a while at least, a pretty steady paradigm of RDD.

And so what we end up with out of sort of this whole context is an increasing difficulty in translating from the respondents that we end up with at the end of the day to our population. Even though as we just discussed during Bob’s talk, we’ve got sort of this, not as much worry as we could potentially have on the lack of a perfect theory linking our sampled population to society at a 9 percent response rate, we seem to empirically be doing okay. But there’s a lot of reason to worry that in this environment sort of the empirics could at some point end up a little bit off and if they do, what are we going to do about it?
And so that brings us in certain ways I think as a discipline to a bunch of opportunities. And in particular I see these opportunities as a lot of these new forms of data we’ve been talking about, things like social media data, things like the mobile phone data that Michael Link was talking about a bit yesterday, the tracking data, marketing data, these different kinds of data sources that we’re gathering ostensibly about individuals though to some extent we don’t really know and trying to attach onto the data that we’re collecting via surveys we’re potentially trying to analyze in their own right. We also have a bunch of new modes of data collection and more sophisticated analytical tools. And particularly we’ve been talking about the use of Bayesian statistics and I’m going to talk a little bit more about that but there are a bunch of new ways that we have to think about linking between the people that we end up with at the end of the day and society.

And the question then becomes one of sort of what model are we going to use to make that link? So the big question that we sort of end up with in this environment I think for people, like the people in this room and the people who are funding the kind of research that most of the people in this room do is the question of sort of can these opportunities that we’re seeing here in some way offset the challenges induced by this increasing difficulty to link between respondents and population. And in particular can we use these new methods for data collection analysis to help us get that kind of understanding whereas there’s something else we really need as a field to make those kinds of links.

So what I’m going to look at in this case and what this particular study is, is taking some consumer file marketing data. In this case it’s data that was purchased from Marketing Systems Group by GFK and the data originally sourced from Experian, Axium and Info U.S.A. in finding out where they come from before that seems to be an impossible task. And these kinds of data, in this case we’re looking at demographics, can be easily purchased even though they end up being somewhat expensive in many cases, they’re readily matched to many of the people that we’re sampling because we have addresses on both the ancillary data and on our respondents frequently and they ostensibly provide a really rich source of individuals, notably for all individuals in our sample, not just for our respondents. So we’re not dealing in this case just with the individuals who responded to our survey, we can now claim to know something about the full set of individuals in the sampling frame.
So if these kinds of data are high quality, we get a bunch of really cool abilities. First of all we can really improve the efficiency of sampling. Right, we can go out and we can target hard to reach groups, we can oversample in particular in Hispanic and young populations and if we do that based on these ancillary data, what we might end up with is a case where we’re actually spending far less money per respondent. We’re getting surveys off cheaper as it were and we’re able to still get our questions in the field in a more efficient way.

Another potential real boon from these kinds of data are that we might be able to get a lot more information on who our non-respondents are and what their features are. If we’re purchasing data and can say something substantive about all the individuals who aren’t responding to our survey, we now have a much better sense of how much nonresponse bias we have, what kind of story that is. And additionally it might allow for corrections for nonresponse and even potentially if we get really good at this stuff and we figure out what sort of the differences are between a general sample and perhaps some kind of biased sample like an online non-probability sample, we might even be able to correct for problems in the sampling frame. So the potentials here I think are enormous if we have big data and get them to work in this one to one matching form with traditional surveyance service.

They’re also as a number of people have mentioned is a long history of trying to connect auxiliary sources of data in various auxiliary sources of data in various forms. And much of our waiting strategy set comes out of this idea that we’re connecting it to the CPS data or to something else. There have also been more specifically, there’s also more specifically an emerging literature on using individual level non-survey data, things like the paradata that Frauke was talking about yesterday to correct for nonresponse. And so there are some tools that are emerging at least that move in this direction and help us think about how to connect these sources of data and use it for various purposes.

But I want to suggest that when we’re going in this direction there are a bunch of questions we have to ask. The first is what are we actually planning on doing with this data. Are we planing on analyzing the data in and of itself which is what the people who for instance are saying well, the tweets seem to match the election we’re doing. Are we looking at it as a way to supplement survey data which is what I’m going to show you in this context. Or are we thinking of potentially other uses off of it, imputations, ways of thing of sort of more complex problems?
The second thing I think we really need to ask are how accurate these data in describing the units they claim to describe. So not how well do they describe the population first. First question we need to really get at is how accurate are they? Is what we find out about from a piece of ancillary information, true, valid, does it match what we find out from other sources of inference, et cetera.

Second, how complete are they? One of the things that I’m going to talk about a fair bit today is actually missingness in ancillary data. Quite frequently we don’t have full ancillary data on every individual. Is that a process that’s relatively ignorable or is it non-ignorable and how much should we worry about the completeness of ancillary data? The third one that I or the fourth one that we really need to address is the model that we’re using to link our data with the world. And Bob was – the whole discussion about use of Bayesian statistics and moving toward sort of thinking about the models linking the individuals that respond to our stuff to society is really about this story. How do we think that the new data that we’re grabbing, whether it be a tweet, whether it be some piece of information we’re buying from a company, whatever else, links between the individuals that we’ve got that information for and their actual state and the world state and that’s a difficult challenge but one that we really as a discipline need to address.

And then finally how do different – how does the model perform across different types of inference? And so even once we’ve gone through ‘em we’ve said, “Okay, we’re comfortable using this data, we understand how they relate to our respondents to the world, we get this process,” we still need to think about whether what we’re concluding is indeed true not just potentially for something like point estimates or the ability to predict the election but also are we interested in how variables relate and does that hold the same on whatever new source of data we’re using or whatever model we’re using to connect these things. Is that true for trends over time? Is it true when we’re thinking about experimental interventions, et cetera? So we have to really seriously consider types of inference when we’re working in this arena.

Okay so when it comes to evaluating consumer file marketing data, as I noted we can sort of have our sample of individuals, we can append the ancillary data to it and what’s really notable is of course this fact that we have, this ancillary data both for our non respondents and for our respondents. And so because we’ve got it for our non respondents we have the potential to use it for some additional correctives. The current project here, first I’m going to
show you some assessments, the correspondence between ancillary data of this sort and self reports. So I’m going to show you the extent to which we end up getting the same and different answers and how different those answers are.

Second I’m going to evaluate a little bit of the nature of missingness in ancillary data and particularly try to assess whether the ancillary data that we purchased in this case is ancillary data that seems to be missing at random in some way that we could understand or potentially non ignorable. And third I’m going to explore whether correctives using ancillary data might give us a better sense of the entire sample, not just our respondents. And so in this case what I’m going to do is use a combination of the self reports and the ancillary data to impute using multiple imputation results for all of the non respondents. And presumably if the ancillary data do a good job at least of mapping the differences between our respondents and the full rest of the sample, this should work even if the ancillary data themselves aren't all that accurate. And so this sort of tells us how well we’re doing in terms of the variables that we’re using and accounting for these discrepancies.

Okay, so the data that we’re going to be using in this case, 25,000 households sampled by GFK from the USPS computerized delivery sequence file. It has over 95 percent coverage. It’s an address based sample. It was recruited via mail in January 2011. Respondents who needed Internet access were provided with Internet access and the results I’m going to be showing you are from the core adult profile which is the first survey they get upon admission to the panel. And so what we end up with at the end of this process is self report data from 4,472 individuals in 2,498 households that were successfully recruited to knowledge panel. So that’s an RR1 of 10 percent. Ten percent of the households that were reached out to were actually recruited into the panel in some way, shape or form. Though we have more individuals from each of those households because GFK allows multiple individuals to be recruited into the panel per household.

The consumer file data that we’re using again is this marketing systems group data. It’s merged with all sampled households and there is a 100 percent match on the addresses. So because the ancillary data, companies do indeed use the same computerized delivery sequence file data, they do have matching results for all of the households so we’re not dealing with a case where we’re not matching here. We’re using the same frame to collect data off both methods.
We produced actually 8 different sets of weights, 4 different types for 2 different levels but the one that I’m going to show you is what I’m going to call the best ancillary match weight. For most intents and purposes this isn't too relevant. All of the sets of weights give the same story but we’re actually biasing these weights toward a match between the ancillary data and the respondents by finding the respondent in each household who most closely matches the age given in the ancillary data. One note about these ancillary data and I’ll talk about them in a moment, is that when they’re purchased from MMG they are purchased on a household level.

And so you end up only with age and education information for this mysterious individual called the head of household. And we don’t quite know what the head of household is but what we did to try to find the head of household in this case is find the individuals who were closest to the age in the ancillary data. So that was our sort of trying to target those individuals set. We also adjusted the weights either to match respondents or to match all sampled individuals. And the respondents are used for assessments of corresponding miss and missingness and all sampled individuals are used for all the analysis I’m going to show you with multiple imputations.

Okay, I’m just going to tell you none of what I just told you actually ends up mattering for any of the analyses. So measures we’re going to use, only six variables in this case. There were not a ton that were purchased here because they end up being somewhat expensive but we’re looking at three household level measures which is whether you own or rent your home. Household income and household size and three individual level measures about this mysterious head of household, marital status, education and age. Okay, first let’s look at correspondents and our basic strategy is to assess the proportion of matches between the ancillary data and self reported data and also to look at how many pieces of data are relatively far off. Here’s basically our comparison.

And so we look at something like this. This is the self report down here. You can see that the ancillary owners tend to match owners a little bit better than ancillary renters tend to match renters but overall when it comes to home ownership we do pretty well here. We have about 90 percent agreement on ancillary home ownership and self reported home ownership. This is not necessarily as good when we look at something like household income and in this case
what I’m showing you is income categories and the difference between them. So right here we’ve taken the income category and subtracted the ancillary data category from the self report category.

And what we see here is only 23 percent of households end up being in the same income category in the ancillary data as they are in the self report data. And fully 44 percent of households are more than one category off. So weren’t just off by a little bit but are actually off by I think the smallest category set was $10,000.00 so by at least $10,000.00 from what they self report so some relatively large discrepancies. And we end up finding similar discrepancies for a lot of our other variables. When we look at household size there’s about 30 percent agreement for the number of individuals in the household and about 33 percent who – or 32 percent who are off by not just individual but two or more.

For marital status interestingly here in the ancillary data more people who are unmarried come across as married than come across as unmarried. So you end up actually getting the wrong answer for most people who self report as unmarried. Overwhelmingly in the ancillary data people seem to come across as married for some reason. But again since these are a black box we don’t quite know why. We end up with pretty decent agreement but really it’s all about the married folk, not about the unmarried folk.

For education we end up seeing again about 40 percent agreement, not miserable but we still have 20 percent of cases that are off by more than one education category. Age, this is the one we bias toward agreeing and we’re going to do it even more by saying if you're within one year because maybe there’s a lag in the amount of time it takes to get the ancillary data to match or something, we get 70 percent but we still have almost 20 percent of our data where no one in the household is within 5 years of the respondent. So there seem to be sort of varying levels of correspondence between the ancillary data and the self report data but considerable discrepancies were pretty much all the variables we’re looking at. If we’re thinking about this as something that’s going to give us the same answer as what we’d get by other methods, probably not going to do so.

Okay, missingness –

_Audience:_ Can I just interrupt quickly? You had a name in both cases? I mean did you check and see whether or not these names –
Josh Pasek: So we did not get a name from the ancillary data. In this case we’re aggregating at the household level so this is all address matched. When we’re dealing with the individuals it’s for the individual we think is most likely to be the head of household.

Audience: You could’ve gotten a name couldn’t you from – when you bought the –

Josh Pasek: Charles is going to have to answer that one ‘cause he’s the one who actually purchased the data ‘cause he’s the one who purchased the data from MSG.

Charles DiSogra: Right, I really appreciate what you’ve done with it Josh. Yeah you could get a name from the listed, list of telephone numbers. You could provide a name with it but we don’t use the name either in the mailing or for analytical purposes but you’re right you could find a name. Now that name may not be as you said, I think the key thing here is that we recruit multiple people per household but you looked for the person that most likely resembled the ancillary information you had on the household. So that would have to be done independent of any name if you were going to do that.

Josh Pasek: Yeah so we’re biasing it as best we can toward a match given the data we have. It’s not could we purchase better consumer file data unquestionably? You know if we wanted to we could probably purchase people’s Social Security numbers or something off of this which is a horrible thing to think about but we can get tons of additional data from these sources. Right now we’re sort of working with the kind of set that a survey organization can reasonably sort of grab and use.

Okay, so second thing we wanted to do was evaluate the nature missingness in the ancillary data. So how often do we not have ancillary data for a particular variable about a household? I told you we had 100 percent match so it wasn’t that we were missing correspondence in the file but what we end up with is actually pretty much a similar story. So I’m going to jump ahead a little bit ‘cause I realize we’re probably a little short on time.

So here’s our missingness by ancillary data variable. We’re missing a little bit less for the household level variables than we are for the individual level variables. There’s no variable for which we’re not missing at least 6 percent of respondents and on ancillary age we are missing 28 percent of households, do not have an ancillary age that was attached to them. Now again we could probably buy other datasets that would have another ancillary age that we could merge on but this is what we end up finding out of
an already aggregated dataset across a number of different variables.

Distribution, it’s not that these are missing for the same people consistently. We end up seeing here is for most – or for about half of people none of the ancillary information’s missing but for some people most of it’s missing. A lot of people are missing one or two variables though.

Okay, if we start looking at it what we end up finding is that missingness appears to be non ignorable by the self report category of the same variable. So if we look at missingness on homeownership, renters are – have a almost a 30 percent chance of missing home ownership. Owners only have a 7 percent chance of missing home ownership. For household income we end up seeing that incomes lower on the income bracket are missing around a little over 10 percent on average while those at the high end of the bracket are missing under 3 percent. For household size we’re missing more information for households that self report as smaller.

We’re also missing far more information for individuals who self report as not married than for married individuals. For education we actually don’t have much differential missingness. It’s not a significant overall though there’s some suggestion that we might actually be missing a little bit more data for our – the people who self report the highest education. And when we look at age we’re overwhelmingly missing ancillary data on younger people. So again – and how well can we predict this, well if we take it all together and we predict the number of missing variables we end up finding a handful of predictors that end up mattering, non owners seem to be more likely to be missing information households with fewer people, unmarried, younger but overall we’re actually not doing a great job of predicting where ancillary data would be missing. So it’s not a process that we seem to understand very well even though it clearly seems to be non ignorable.

Okay so ancillary data as something we might rely on don’t look that great given these analysis but maybe essentially if we go at it with a Bayesian brute force tool we can get around this anyway which is one of the possibilities that we’re kind of interested in especially if we want to understand our non respondents. Maybe the cross cutting nature of these ancillary still tell us something substantive. And so what we did here is we imputed the distribution of self reports for all sampled individuals, not just our respondents based on the ancillary data and we wanted to see sort
of if that was a substantive improvement over the self reports of respondents alone and how much. So here’s the basic procedure. We’re imputing it for the whole sample and we’re going to compare the imputations, the raw self reports, unweighted mind you, and the ancillary values to the CPS. We used a bunch of extra measures in the imputations to the extent possible just because the more you shove into an imputation the better it should do and there are lots of reasons that that tends to be the case as long as you think that the measures that you're adding to the imputation aren't going to be biased.

Okay, I’m going to show you results from 100 imputations using multiple imputation via chained equations which is one of sort of the best you shove the data in there, it spits out a result imputation programs in the market and then we’re producing these point estimates. So here’s – that was not good. Come back. Okay, so here’s what I’m going to show you for each of these and I apologize, these are a little funny to interpret. But this green line here is the current population survey estimate, the dashed green line. The red line is our raw data estimate from GFK so this is unweighted not even dealing with the demographic issues that we usually have.

We’re looking at demographics, that’s the hardest thing to do. This box plotting here is the range of our multiple imputations. So this tells us how much our imputations range across the 100 imputations and the blue line is our ancillary estimate. So what we see for instance from home ownership which is one of the variables which we did pretty well on in the first time is that we do about equally well using the raw GFK estimate as we did with the imputed data but that the ancillary estimate was actually a decent bit off.

I have a whole bunch of these but what you can basically see looking across these stories and I’m going to show them to you reasonably quickly 'cause I have one for each set and I’m happy to sort of walk through them if we need be, is that by in large across these, the raw GFK and the imputations tend to be the closest ones to the CPS estimate. The imputations end up doing a little bit better on average and the ancillary data themselves if treated as an estimate of the population always generally seem to be quite a bit further off. So we have that with regard to household size. We have that same story again with regard to marital status. Here the imputations seems to do pretty well at getting near the CPS estimate. When we look at education again sometimes the imputation seem to be really close, other times they seem to be
further off. Age, the same story. And so across all of these, sorry, Bob?

_Audience:_ [Inaudible question]

_Josh Pasek:_ Raw in this case is – it’s not weighted to the population, it’s just taking the self reports as they are.

_Audience:_ [Inaudible question]

_Josh Pasek:_ It’s not weighted to the population; it’s just taking the self reports as they are.

_Audience:_ [Inaudible question]

_Josh Pasek:_ It’s adjusted for the base weight. There was a slight differential probability of sampling in that GFK actually used some of the ancillary data on Hispanics and young people to supplement its initial pool and so those individuals are down weighted. And then we also weighted to the household level so those two weights are on the data that I’m calling raw right now but nothing that’s a specific attempt to match the population, just an attempt to see who sampled.

Okay, so across all of these what we end up with is a story where the imputations tended to perform the best by not by an enormous margin. The raw estimates were sometimes a fair bit off but never enormously off. The ancillary estimates were pretty far off, particularly in the case of marital status but almost always much further than across the other variables. And so on average we end up seeing that imputations performed the best though the difference between those imputations and the CPS is still more than half of what we ended up in the raw data with no correctives to try to address differential and non response across groups at all. So how you interpret this I think is an open question but it doesn’t look like with six variables we’re really eliminating most of the error anything of that sort.

If you want to look at the numbers for this, particularly look at the total numbers. You can see the raw self reports on average, again, not corrected for the population differed by about five percentage points from the census. The imputed data differed by about three percentage points on average. So we get rid of some of the error but not most of that difference. And I should note that the self reported variables are almost identical to the census and also the household level sampling and how proxies are dealt with on the
household level is also very similar to the way the CPS collects its data. So better again but not by an enormous amount.

So what do we take out of this? So first of all estimates from the raw self report data weren't all that far off. The imputations based on the ancillary data did eliminate some portion of those errors that were in the self reports and that the ancillary data themselves are just not something you’d want to rely on as an estimate of the population by any means. Taking all these analyses together you know, we find that there are frequently discrepancies between the estimates of the ancillary data and self reports. The missing ancillary data seems to be systematic and it appears to be non ignorable. And the standard base imputation algorithms don’t fully correct for the biases even though they do seem to improve things somewhat.

So what’s this suggest about using this kind of data and where we are at the moment? Well they may be useful for reaching really hard to reach populations. I mean it was definitely true that there were more younger people that were identified as younger people in the ancillary data than young people that were identified as older people. So maybe on that level it might provide somewhat of an improvement to efficiency but again there’s sort of a distinct bias variance tradeoff with any given variable that you want to pay attention to in that regard. The ancillary data don’t seem particularly efficient in correcting for non response and it’s unlikely that it’d be possible to use these data given the sort of moderate improvement we saw in non response to correct for something like a bad sampling frame.

So we’re probably not in that league, at least with the number of variables we have here. So the question then is sort of what went wrong and the problem as we’ve discussed pretty heavily is we really don’t – can’t know that the ancillary data themselves are a black box. We don’t know what’s actually measured, what’s imputed, how they’re linked between various data sources. This is the critical problem we’re dealing with here. And so because we’re trying to peer in one – the out coming end of the black box and not the ingoing end, we don’t really know where these biases are coming from and what we end up measuring by the end of this process.

So moving forward from here, you know I think there are a lot of reasons that we really do want to encourage good ancillary and auxiliary data of these types as a substantive improvement to survey sampling and as additional things we could work with but
the kinds of demographic ancillary data that we used in this study weren't sufficient for a lot of those purposes. They didn’t seem in general to be high enough quality and we probably at the quality they were would have needed a ton more of them to correct substantively for non response and for issues like that.

With the current data we probably do want to know more concretely how they compare with things like traditional weighting techniques. We want to be able to think about using a larger set of data so maybe with a lot more measures we can sort of brute force our way out of this problem and maybe we can think about linking even more types of data to do better or going for additional data sources. But probably the right answer is the same one that Bob was alluding to which is we need to think about how to get data we can trust and evaluate. And to do that we need more transparent ancillary data.

The process of linking data sources to one another needs to be more systematically addressed. It’s something we have to deal with if we want to start incorporating these additional sources. And maybe we really want to be focusing on an in-house option, on something that we can do ourselves that means we’re not purchasing these data from corporations that aren’t going to tell us quite how they work and start thinking about well, can academicians build a dataset that really does start thinking about these links that starts putting these pieces together. And I think the NSF can play a pivotal role in putting something like that together. And so that would be sort of my strong suggestion out of here.

Just to conclude very briefly, I think when we’re thinking about these additional sources of data I want to stress these as important questions that we really need to think concretely about what we’re doing with it, how accurate they are, how complete they are, the models that we’re thinking about using and how those models perform for different types of inference and that strikes me as the question set going forward. Thank you.

Tom Smith: Tom Smith. There was a NSF sponsored workshop last year. Tim, Bob and I were involved in it and which was strongly supportive of using auxiliary data and paradata, many of the things that mentioned here for that. Some of the specifics of your particular findings and some of the things that I found in my own research do show the quite varied differences between trying to use auxiliary data. For example you said there was 100 percent match across addresses. We found it very difficult to match across addresses when it was apartment buildings because many databases do not
specify the unit in an apartment address which made it very
difficult to take an address for about the 20 percent of addresses in
our sample which were that – not that all them couldn’t be matched
but it was very challenging to do that and they certainly couldn’t
all be reliably matched there.

So that’s one thing which you got more pos results. On the other
hand the main database that we use automatically gave us the
names of all adults in the database and overwhelming, although not
100 percent of the time the ages of those adults so it’d be very
much easier to match to a specific individual from a survey to
those households in those sets of auxiliary data. So I’m just saying
that this is a new and evolving field and there’s going to be a lot of
variation results depending upon the different databases that are
used and the particular way they are thing.

The last comment I’ll make is about income. I had in-depth
collection with a number of those proprietors of the black
boxes, particularly about income. And it is my general conclusion
based upon what they said is the vast majority of income data is
imputed would even be too strong of a word, guessed I would say.
Very few of them actually have a stated income and that’s why I
think you're getting some enormous variation on income particular
as opposed to other things like age and marital status. It may not
be right but usually they got some kind of a concrete report about
some of those things. But for income usually it’s something, they
had an occupation, they had some kind of credit card spending
thing, that they just had some bit of evidence from which they
came up with an income.

Josh Pasek: Yeah, I mean it’s clear these vary wildly and the question then is
okay, as practitioners at the end of this, how do we pry that box
open to get enough of a sense of what we can trust and what was
created via system that we’d all shudder at. And there clearly is a
lot of data that we’d all shudder at that’s coming out of these
datasets but we have no idea what they are. And I think income
may well be a good case of that. Bob?

Robert Groves: This is Bob Groves. Real interesting stuff. I’m with Tom on one
thing and let me put a plug: I mean stay tuned for a report that
should come out of census soon that is essentially purchase four of
these major datasets and then passed the records by the 2010
census. The address match issue is an issue and it really is and it
looks like your vendor blocked you from the problems they had at
matching.
Josh Pasek: Almost certainly.

Robert Groves: Yeah by giving you DSF things. I also know DSF has coverage problems right? Not everyone gets mail at their house. And those addresses are problematic. DSF has unit problems attached at the apartment number. Okay? But that should be useful because the match algorithms, there’s a little sensitivity analysis on the matching which I think is important. It doesn’t have a bunch of rich variables that we’re interested in.

Secondly it seems like you're one step away from what the real question is and that is, are the variables that were the purpose of this survey affected by all this stuff? You just did demographic comparisons to CPS so it seems like you could do sensitivity analysis. I assume this survey wasn’t done to estimate income.

Josh Pasek: Well it actually kind of was. This was to get the demographic core profile for people joining knowledge panel. So unfortunately it’s not a survey that’s really about –

Robert Groves: I see, you don’t have other variables.

Josh Pasek: I think there are additional ones I could get from GFK and you know they’re – we can start to estimate some of those and what they are but this is not a field survey done for a traditional purpose. It’s to join a panel.

Robert Groves: This is by way of standards, it seems like if we’re going to do this stuff and you demonstrate that there are biases on variables for which we have population parameters already, that’s not very useful because we’ll adjust away those things through survey weighting. So we all ought to force ourselves to look at other variables. And then you won't have it as CPS comparisons but at least you could do sensitivity analysis pre and post adjustment by what you –

Josh Pasek: I think that’s a good call and I think that’s where this needs to move as sort of a first step it seemed important to just get some quality control questions on here and to see how well are we doing at estimating the stuff we do actually know with stuff we know. And the answer was not all that well actually. Andy?

Andy Peytchev: Andy Peytchev. Well first of all I think this holds great promise but having said that, in addition to the linkage problem, and in addition to problems about the quality of the data for example, the income being highly imputed and a black box at that. There is also
the problem of what are the variables in that database and how useful are they for the variables in the survey? So Ragunathan and I have actually a grant to evaluate exactly that kind of design where we took a survey that some tobacco use and we bought Experian data but those are their Experian data that has a package on tobacco use. And the results I guess the very short answer to the entire exercise is that in addition to all these major problems, the biggest problem is the percentage of records that you could actually link.

Now this is a telephone survey but the percentage, the linkage rate for something like income is very high. That’s great but for something like tobacco use it’s abysmal, it’s in the single digits. So I guess while the methodology holds promise, probably the biggest impact would be finding databases that would have the relevant data that we need for our surveys. It would be highly correlated to the survey variables and managing to clean those up for a – to be fit for a survey use.

Josh Pasek: Yeah I mean I think the answer we’re all pointing toward is this has to be a pretty big agenda of thinking about it, of really making it a problem for social scientists, for statisticians, for computer scientists and not just for the corporate sector that opens this up in a way that we can start to get a better sense of what’s going on.

Audience: Quick question, are the averages from the ancillary data, are those from the link data or from the overall profile? Because we often, the overall profile, if you were to buy the overall profile from these providers often differs from a sample. What we find is when we take small samples the data from these databases varies widely but the aggregate profile tends to settle down.

Josh Pasek: So I, we used the link data on the sample when we just you know, took the entire sampling frame and looked at the average of the ancillary data over that.

Charles DiSogra: This is Charles. I just want to say first of all about since when I bought the sample that we’re talking about I excluded the drop points so that maybe address match easier and also we excluded post office boxes that were not the only way to get mail. So we did kind of maximize what we got from the addresses in that way. The other thing is that Josh here just took advantage of the opportunity that we had both the profile data on these individual self report and we had the ancillary information. The ancillary information was only purchased to do sampling stratification, to improve the probability of reaching certain groups and we had previous years’
data where we stratified based on census block information and we had ancillary information to see how predictive they would be.

And so we only picked age and being Hispanic as the ones we used for stratification. But we always continued to buy the ancillary information and just to see how well it did correspond with the self reports of individuals once they actually profile. And it was that rich data source that Josh and his team took advantage of.

**Josh Pasek:** So the variables were sufficiently different between the two in terms of how they were measured that I decided not to use Hispanic. The other I decided not to use was presence of telephone. The two of those were measured so differently in the ancillary data that I didn’t want to do a direct comparison between them even though I did use them as part of the imputation process.

**Robin Gentry:** Robin Gentry with Arbitron. We’re using something very similar to what GFK does. We’re looking towards trying to stratify by this information. The other interesting area that I think there deserves to be some more research on is how you can use this ancillary data to improve response rates by knowing more about the sample prior to sending materials out to them, being able to target things in the appropriate language, things of that nature. The question I have is within those sub groups where you for instance can get Hispanics or perhaps young adults, how good is the coverage and what are the biases between the sample, the Hispanics you can get and those you can’t get if you're doing it based on for instance surname matching, what sort of biases are there for certain countries of origin and other things of that nature?

**Josh Pasek:** Yeah that’s one we haven't done yet. You know this is very much still a project in progress though this part of it’s relatively complete. I do want to do a fair bit of okay, if you oversample on this particular group, what are the bias variance tradeoffs and how do we understand them? We’ve been oversampling various different groups for a long time. There actually are shockingly few studies of what our bias variance tradeoffs are on a lot of these oversamples and what they can do. So I think that’s an important direction to think about here.

**Robin Gentry:** And I think also that the area of efficiency is interesting but what we found is that missingness is probably one of the most predictive variables essentially. Within the missing sample you have much higher numbers of minorities, et cetera. So how can we figure out the most efficient way – for instance we’re trying to actually under sample older people using this data 'cause the data is very good for
older people but that doesn’t help with efficiency. It at least helps with representation.

Josh Pasek:

You know there are a bunch of different techniques you could use out of the literature right? You could go for something like trying to produce a regression, doing matching, algorithm stuff like that to try to figure out the individuals propensity of responding within various different groups and across various different groups and how they’re going to – you know and what their actual demographics will end up being. The problem is if you can’t predict their actual demographics particularly well from this data, then we’re in a little bit of trouble because we’re going to have a much harder time using this data to do the kinds of things we’d love to use them to do like more targeted sampling, like correcting our frames, like dealing with nonresponse more efficiently and potentially like thinking about this distinction between our sample and our respondents.

Bob Belli:

Bob Belli, University of Nebraska. One thing is really sort of on my mind as someone who’s interested in measurement error and data quality in terms of these linkage issues and I don't know if this is a question for you or a question for people really interested in linkages. And that is just a real basic and simple question that we know that questions and question wording impacts the kinds of answers people provide. And hence when you’re looking at linking various datasets and you have no idea with regard to or very little idea with regard to what the question is, although you may have a sense that these different questions are measuring somewhat the same thing, it could be quite a ways off and apart.

It seems to me for example that some of the data that you looked at, you’re being able to get better levels of agreement with such things as owning or renting a house but when the question became more complex such as income, things really started to diverge rather dramatically. And we can imagine all sorts of different kinds of questions that we’re interested in. Just one example would be unpaid labor, household work and gender differences there. And with regard to my experience in the PSID, what they began to see was a vast increase in one of the years on that panel in which men were engaging in more unpaid labor. Well the question changed. That’s the reason why.

And so when you're beginning to try to look at these linkages and don’t really know the nature of what the question is to begin with, how do you propose to have success? I mean all the statistical manipulations I really don’t understand myself. It seems almost
like a magical process. Maybe there’s something there that I’m missing. But that seems to be a real problem that no one really is seriously addressing.

Josh Pasek: I think that is the issue, the fact that we can buy this data doesn’t get us over the okay, what are the linkages? What are the gaps? What are we describing with this data? Are we describing public opinion the way we traditionally think of it or are we describing something very different? I’m not quite sure.

You know I have done a number of things sort of thinking about and being on panels with regard to thinking about social media data in this regard and when you think about social media data for somebody to actually tweet something is a completely different cognitive process from for somebody to report when asked a question about that same thing. And so imagining that you’d get reports that are even remotely similar is kind of surprising but what we end up at the end of this whole process is there might well be social media data in what I end up with here because I know that Axiom uses social media data but I don’t quite know how. There might well be various different sources and I can’t disaggregate what’s going on. There’s no way given these kinds of data at the end that I’m going to be able to sorta work backwards to figure out you know how these were put together. Basically all we know is we don’t really match.

And so the story that we don’t really match I think gives us an imperative that if we want to think about using this data we have to systematically move toward a serious understanding of it and that serious understanding probably means that academicians have to get into the big data business in an actually collecting and linking way, not just in a purchasing from companies way.

Jon Krosnick: This is Jon Krosnick. I’ll exert the chair’s prerogative to end this session with question for Bob and a question for Stanley just to clarify what you guys are both saying. So if I understand correctly Josh, you're comparing data you purchased with answers to questions that GFK administered during the profile questionnaire.

Josh Pasek: Yeah.

Jon Krosnick: So there is question wording of the GFK question but there is no wording of the supplied question –

Josh Pasek: Right. All you get is married, not married.
Jon Krosnick: So your point Bob I think if I take it right is that GFK could’ve gotten either a closer match or a farther away match depending upon what wording they used in their profile questionnaire and there is no right wording given that there – I mean we can match wording if we’re trying to match a survey but since we’re not trying to match a survey there is no right wording, is that?

Bob Belli: I was trying to avoid the more philosophical issues associated what accuracy means. That’s a probably a even more of a – and it’s probably related in some sense with regard to even if we’re talking about something as – something like objective data. You know we’re making certain assumptions that there is an objective reality out there and to some extent I think that in terms of asking people not they're working or not there probably is some sense of objective reality of working for pay or not working for pay even though there are as you know, based on different scenarios, different judgments as to what constitutes work and what doesn’t constitute work. People have to make judgments on that. And there are various ways of asking questions in which you may be able to tailor those questions in a framework in which however an economist wants to define work may be closer to that economic definition but there’s always going to be some mismatch there.

Having said that with regard the big issue and even more detailed issue has to get a better understanding that the framework of how questions are asked are going to – can potentially lead to vastly different answers and it grows in terms of its complexity from one moves for more objective data to more subjective data such as whether or not as you looked at recently Jon whether or not one has racist feelings or racist attitudes. The manner in which one approaches asking these kinds of questions is critical to the answers that are derived and the inferences that can be derived from those answers. And if you are linking data that you really don’t have a very good understanding even of that – you know of the complexities involved, I just don’t see how you're able to gain any sense of accuracy out of those linkages. That’s the point I’m just trying to drive.

Jon Krosnick: Thank you. Stanley?

Stanley Presser: I was trying to think about the implications of the point you were making. So let’s say they had asked for the names of the people, then they would have found in some fraction of the respondent households that the names didn’t match. So then they would say okay, so we actually don’t have ancillary data on these households at all but they wouldn’t know among the non responding
households how many of those households the ancillary data don’t actually fit those. So I wasn’t sure other than losing a bunch of cases what would the implications be?

*Bob Belli:* All I was really saying is this has a massive implication potentially for the discrepancies that Josh observed when he was looking at how well the ancillary data performed relative to self reports because in some of these households these are entirely different people reporting.

*Josh Pasek:* I mean we did what we could in terms of creating different weighting systems for that though.

*Audience:* No, no, I meant I just misunderstood because I know, I thought this was kind of a primary data collection in which the names would’ve been available. I didn’t realize this a kinda secondary data analysis that this frame was initially bought for another purpose that doesn’t have the names and I understand completely but I misunderstood initially 'cause I realized that in principle those names were obtainable initially. That’s all.