A METHODOLOGY FOR REDUCING RESPONDENT DUPLICATION AND IMPERSONATION IN SAMPLES OF HIDDEN POPULATIONS*

Douglas D. Heckathorn

Cornell University

Robert S. Broadhead

University of Connecticut

Boris Sergeyev

University of Toronto

Forthcoming: Journal of Drug Issues, Spring issue, 2001

*Paper presented at the annual meeting of the ***American Sociological Association*** in Toronto, August, 1997. We thank Robert Mills and Heather Madray for their help in analyzing the data presented in the paper, and Kathleen O'Riley and Brendan Walsh for their help in gathering the data. Finally, we thank the *National Institute on Drug Abuse* (Grant #R01 DA08014) and the Centers for Disease Control (Grant #U62/CCU114816-01) for its support of this research.

Abstract

A dilemma arises for researchers who sample hidden populations, such as injection drug users (IDUs), and use financial incentives to recruit respondents. To prevent respondent duplication (a subject participates in a study multiple times by using different identities) and respondent impersonation (a subject assumes the identity of other respondents), researchers must confirm their subjects' identities. Documentation, however, introduces sampling bias against those who lack such identification, or who wish to remain anonymous. Definitive forms of identification like photography and fingerprints introduce a bias against the more distrustful members of the population, and scanner-based biometrics can be expensive. Most research projects therefore rely on staff to recognize former respondents, but staff turnover and a large number of respondents compromise accuracy. We describe and assess quantitatively the accuracy of a method for subject identification based on a statistical principle, the *interchangeability of indicators*, in which multiple weak indicators combine to form a stronger aggregate measure. The analysis shows that observable indicators of identity (scars, birthmarks, tattoos, eye color, ethnicity, and gender) and five biometric measures (height, forearm lengths, and wrist widths) provide the basis for a reliable and easily administered method for subject identification.

Introduction

AIDS has accentuated the problem of accessing members of hidden populations for research and intervention purposes, especially in the case of injection drug users (IDUs). Hidden populations have three characteristics: no sampling frame is available; their members are objects of hate and stigmatization within the larger society; and their members are distrustful and work hard to avoid identification.

The research literature regarding IDUs has focused primarily on improving ways to reach and recruit them into prevention services (Wiebel, 1988; Broadhead and Heckathorn, 1994; Cunningham et al., 1996; Heckathorn et al., 1999), drawing more diversified or representative samples (Biernacki and Waldorf, 1981; Watters and Biernacki, 1989; Spreen, 1992; Heckathorn, 1997), and estimating the size and composition of IDU communities within municipalities (Hser and Anglin, 1993; Frank and Snijder, 1994).

This paper focuses on a neglected methodological problem and source of sampling bias: respondent duplication and impersonation. Respondent duplication occurs when a respondent participates multiple times in a study by using different identities. Respondent impersonation occurs when a respondent participates in a study by assuming the identity of other respondents.

The difficulties that research and intervention projects face in keeping track of hundreds of IDU respondents should not be underestimated. In many municipalities, especially large urban centers where numerous drug-related research projects can be operating simultaneously—and have been for years—researchers commonly speak of "professional research subjects" who are continually cycling through many projects, or through a single project at its different intake sites. The second author's extensive field study of a large AIDS prevention project for IDUs in San Francisco (Broadhead and Fox, 1990; Broadhead and Margolis, 1993) included many IDUs who described with amusement being interviewed by the same project several times, using different names at any one of its many intake locations around town. Although the financial rewards for a single interview are generally small—$10 to $50 depending on the research project and the demands made upon respondents—they can add up if a subject gives many interviews over several days or weeks. Interview staff also speak of veteran respondents who have learned what responses will substantially shorten an interview from, for example, 11/2 hours to 20 minutes, by claiming to have no sexual partners and to use only a single type of drug. The word on the street about answering most questions is, "Just say no—or none."

The problem of respondent duplication and impersonation has also affected funding opportunities for researchers. The second author has participated on many NIH review panels over the past five years in which community-wide epidemiological or intervention drug studies in major urban areas were scored at a disadvantage because the investigators did not explain—or could not explain convincingly—how they were going to prevent respondent duplication or impersonation in a proposed study.

To better ascertain the scope of the problem, we emailed and received responses from ten well-known researchers who were principal investigators of large, multiyear community intervention projects for drug injectors. We asked them, “In your experience, how common is the problem of ‘subject duplication’ in community surveys of IDUs?” and “Are you aware of any empirical studies of the problem?” None knew of any published studies, but all ten noted that subject duplication was common. Five of the researchers began their email responses with comments such as, “This is a great issue to explore,” and “This is a good research question.” Most of the researchers estimated that subject duplication involves 3 percent to 5 percent of their subjects; the following comment comes from a Chicago researcher:

> This is difficult to quantify given that we don’t catch those who are successful … If I had to guess, I’d say that duplicate enrollments in past studies constitute well under 5 percent of our samples, and that we eventually identify the majority of them. However, I’m also willing to say that we could have been “chumped” to a greater degree than I recognize.

A researcher in Denver noted,

> We discovered a number of duplications. During this project, and in all other projects, we tried very hard to avoid this and “caught” a number of individuals trying to get into the study twice. We found 30 confirmed duplicates out of our cooperative agreement cohort of approximately 1,000 subjects. We also turned down a number of folks who tried to come more than once.

Finally, a Seattle researcher emphasized,

> In our experience, there is an important and common problem of duplication … Our observations were that there were some people who just enjoyed enrolling twice or (in

one case) three times. It was thrilling for them to get away with it. There is fairly strong motivation to try it especially if you have enrolled and can visualize how easy it would be to do it again.

In this paper, we present a methodology for reducing significantly the possibilities of respondent duplication and impersonation. The Identification and Reward Information System (IRIS) was developed as part of an AIDS prevention study for IDUs called the Eastern Connecticut Health Outreach (ECHO) Project. IRIS is a computer-based information-processing system for recording and retrieving a combination of non-sensitive biometric measurements for each respondent (height, length of forearms and wrist widths), and visible physical characteristics (scars, birthmarks, tattoos, eye color, ethnicity and sex). Taken together, the accuracy of these multiple indicators for identifying any given respondent is high, yet the subject is not required to reveal his or her real name, social security number, address, or provide any other form of positive identification. With the use of a lap-top computer, IRIS is also fast and easy to use under field conditions, as in community storefronts where respondents are screened for initial and follow-up interviews, or when they return for subsequent services or to receive respondent fees. To protect subjects' confidentiality, the program is password protected, and sensitive information is written to the computer's hard disk in encrypted form.

Background

The incentive for respondents to engage in respondent duplication and impersonation is financial. It is customary in AIDS prevention research to compensate respondents for the sometimes-considerable time required to participate in a study. Interviews frequently last one to three hours, and corresponding fees range from $20 to $50. Specific sampling methods can increase the incentive and thus aggravate the problem. For example, in targeted sampling (Watters and Biernacki, 1989), interviewers fan out into targeted areas to conduct interviews over the course of several days. A respondent can move from one interview site to another, collecting respondent fees along the way. In a peer-driven intervention (PDI), which the ECHO Project utilized, respondents are paid financial rewards for initial and follow-up interviews, and for recruiting three or more respondents into the study and disseminating risk reduction information in the community (Broadhead et al., 1995; Broadhead et al., 1998). Respondents return to the project site days or weeks after their interviews to be paid for their education and recruitment efforts. A PDI therefore gives respondents reason to attempt to impersonate others in order to collect their recruitment and education rewards.

In research that involves sampling hidden populations, such as injection drug users, any requirement that respondents confirm their identity with documentation, such as a driver's license or social security card, introduces a bias against those who lack such identification, or who do not wish to divulge it. This bias tends to affect the most dispossessed part of the population, including the homeless, the geographically unstable and non-legal residents. Definitive forms of identification can be generated through fingerprints or photography, but these introduce a bias against the more secretive, paranoid or distrustful members of populations who seek to remain hidden for valid reasons. Given that disclosure of one's membership in a hidden population can cause the loss of a job, estrangement from family and even imprisonment, such a concern is reasonable. Therefore, most research projects involving IDUs use less-than-definitive means of identification.

A common approach is to rely on outreach workers, interviewers and other project staff members to recognize respondents. Reliance on facial recognition reduces the extent to which the sample is biased by the selective exclusion of respondents, but it has its problems. First, the accuracy of this approach degrades as the number of respondents and duration of the study increase. For example, the ECHO Project administered interviews with 896 IDUs over four years. The number of respondents was too large for any one individual, or even a team, to keep straight. Second, staff turnover tends to be frequent in AIDS prevention research. When an interviewer departs, the ability to recognize respondents by face is compromised. Third, in large projects with many interviewers, respondents can participate more than once without coming into contact with the same personnel.

Another approach is to give respondents code names or numbers to use to identify themselves, but this is unreliable because respondents frequently forget their assigned codes. Respondents can also collect different names and numbers that enable them to participate multiple times. Still other projects rely on a combination of observable characteristics, such as ethnicity, gender, approximate age and physical characteristics, such as scars and tattoos. This approach is also limited because the specificity of demographic variables is limited, and many respondents lack unusual traits.

Biometric Measures and Respondent Identification

Means for respondent identification can be arranged along a continuum from *hard measures* that are highly reliable, such as legally valid picture IDs or photographs, to *soft measures* of lesser accuracy, such as facial recognition by staff members. New hard measures are becoming available, including digital scanners that base identification on palm prints and fingerprints. However, given the association between fingerprinting and the criminal justice system, these may prove even more threatening to respondents than photographs. Scanners that base identification on the iris of the eye may be more promising, but at present they are prohibitively expensive. Facial recognition scanners are under development, but for now their accuracy is limited.

Thus the dilemma remains: definitive means of respondent identification are perceived by members of hidden populations as threatening, and the non-threatening means are inaccurate. An ideal way of identifying respondents would combine the accuracy of hard measures with the non-threatening character of soft measures, and it would be based on information that program staffers could quickly and reliably gather and share among themselves.

The key to constructing a method that satisfies the above criteria is the statistical principle known as the *interchangeability of indicators,* which shows that *multiple indicators of low accuracy can produce in combination a highly accurate indictor*. This is the principle upon which scaling is based. Any single scale item is less reliable than combined items. Similarly, an identification scale can be constructed using multiple indicators, each of which is both durable over time and sufficiently lacking in accuracy to be non-threatening to the members of the hidden population.

The use of biometric measurements provides a potential resolution to the dilemma. Certain measurements are reliable and provide reasonable specificity yet lack the threatening associations of fingerprints, photographs or other forms of positive identification. The most reliable are skeletal, such as length of forearms, height, width of ankles and wrists and hat size. For adults, these measurements do not vary over many years, even if a person gains or loses considerable weight or radically changes appearance.

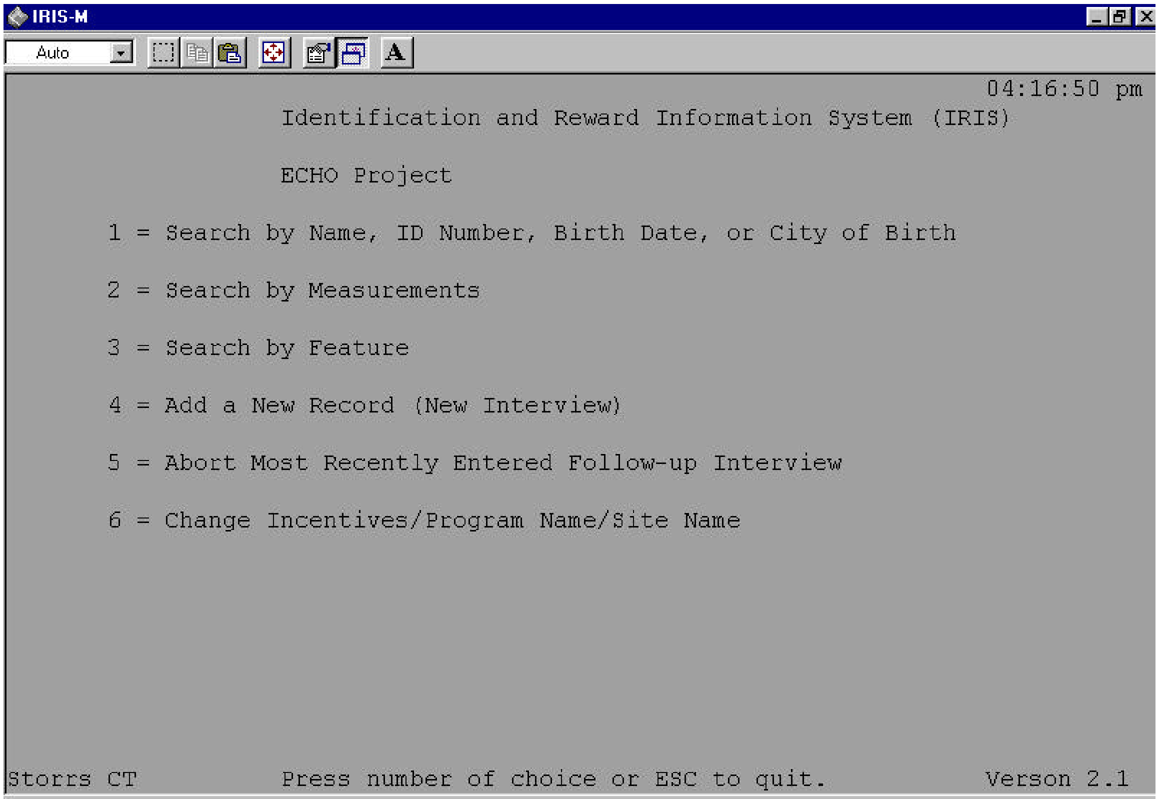
IRIS relies on measurements of respondents' wrists and forearms. Because bilateral asymmetry is common, measuring both forearms and wrists increases the overall accuracy of the measures. IRIS also relies on respondents' height (within three inches) and visible characteristics, such as scars, birthmarks, and tattoos on their forearms, hands, neck and face. These data establish "target values" for each respondent that are discriminating when considered together. The analysis below demonstrates that in combination with observable personal characteristics, biometric measures can make a significant contribution to solving the problems of respondent duplication and impersonation.

The Use of IRIS in the ECHO Project

In relying on a PDI, the ECHO Project paid IDUs nominal monetary rewards for an initial interview and three six-month follow-up interviews ($20 and $30 respectively). Following each interview, respondents were offered the opportunity to earn further rewards by educating three peers in the community and recruiting them to the project's storefront for interviews. Each respondent was given three recruitment coupons and informed that he or she would receive $10 for each IDU peer recruited into the project. In addition, he or she would receive up to $10 for educating each recruit, as measured by a brief knowledge test administered to the recruit before the initial or follow-up interview. Thus, each IDU respondent who recruited and educated three peers in the community was eligible to earn up to $60 for his or her efforts, and each was given the opportunity to do so again at six-month intervals. In turn, each IDU who was recruited was also offered three coupons to recruit still more peers. With this "respondent-driven sampling" mechanism, the number of respondents recruited by the PDI expands geometrically over time (Heckathorn 1997).

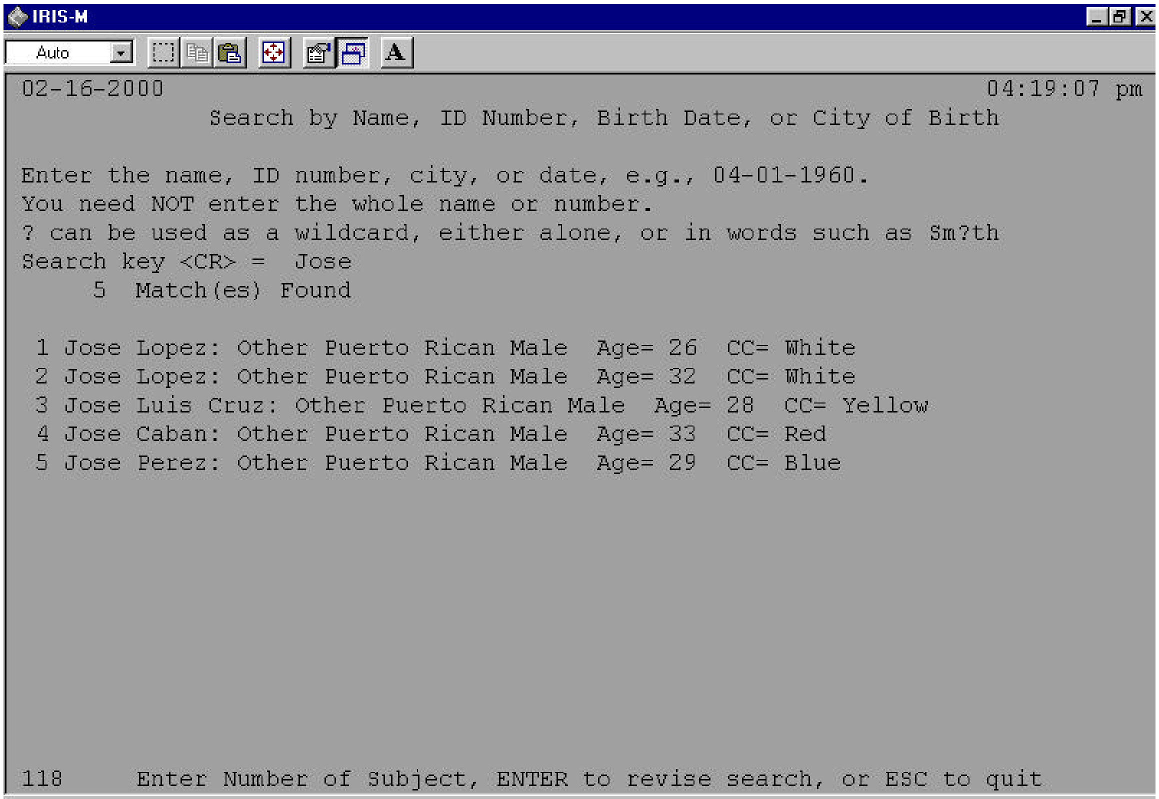
When IDU respondents arrived at the ECHO Project storefront for their appointments, a Health Educator greeted each one individually and ushered the respondent to a private room where he or she was screened for injection drug use. After gaining the respondent's informed consent to continue, the Health Educator recorded demographic information, height and color of hair and eyes, and all visible physical characteristics on forearms, wrists, hands, neck, chest and face. The Health Educator then measured and recorded the length of the respondent's forearms from elbow to tip of index finger using a straight-edged ruler, and the width of the wrist bones using digital calipers. At this point, if the Health Educator wanted to check whether the respondent had been interviewed before by the ECHO project, the Health Educator could use IRIS to run a series of searches comparing the respondent's features with those of all respondents in the computer database. Figure 1 presents the IRIS search options on the computer-screen available to the Health Educator:

Figure 1: IRIS Screen



The Health Educator can search the database by name, respondent number, date and city of birth, biometric measurements and physical features, such as scars and tattoos. For example, she may wish to search the database for a new respondent who is Puerto Rican, male, 5’2” and age 41 and asks to be called José. The respondent has a tattoo on his upper right arm that reads “Debbie.” The following search on the name “José” of an ECHO database consisting of 175 respondents in one eastern Connecticut town yields the following:[1]

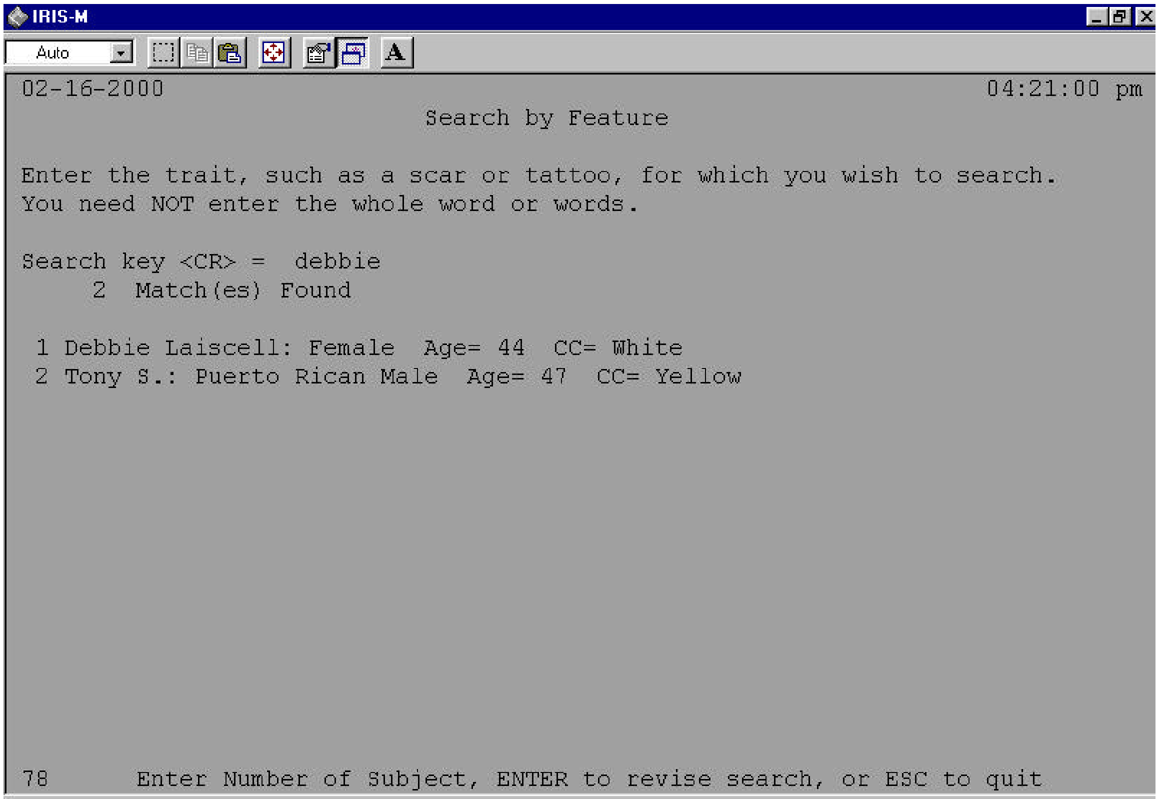
Figure 2: IRIS Screen



(In this and all subsequent examples, *respondents' names and personal identifiers were altered to protect their confidentiality.*) Note that five former respondents gave "José" as their name. The Health Educator may well wonder whether the man sitting before her is one of these respondents. By clicking on José #1 and reviewing his Case Summary, she learns that, with respect to height and tattoos, the present respondent is definitely not "José Lopez." IRIS enables her to scan the Case Summaries of all remaining "José" respondents, and in less than 30 seconds, she confirms that based on differences in height and tattoos, the present José is not one of those respondents, either.

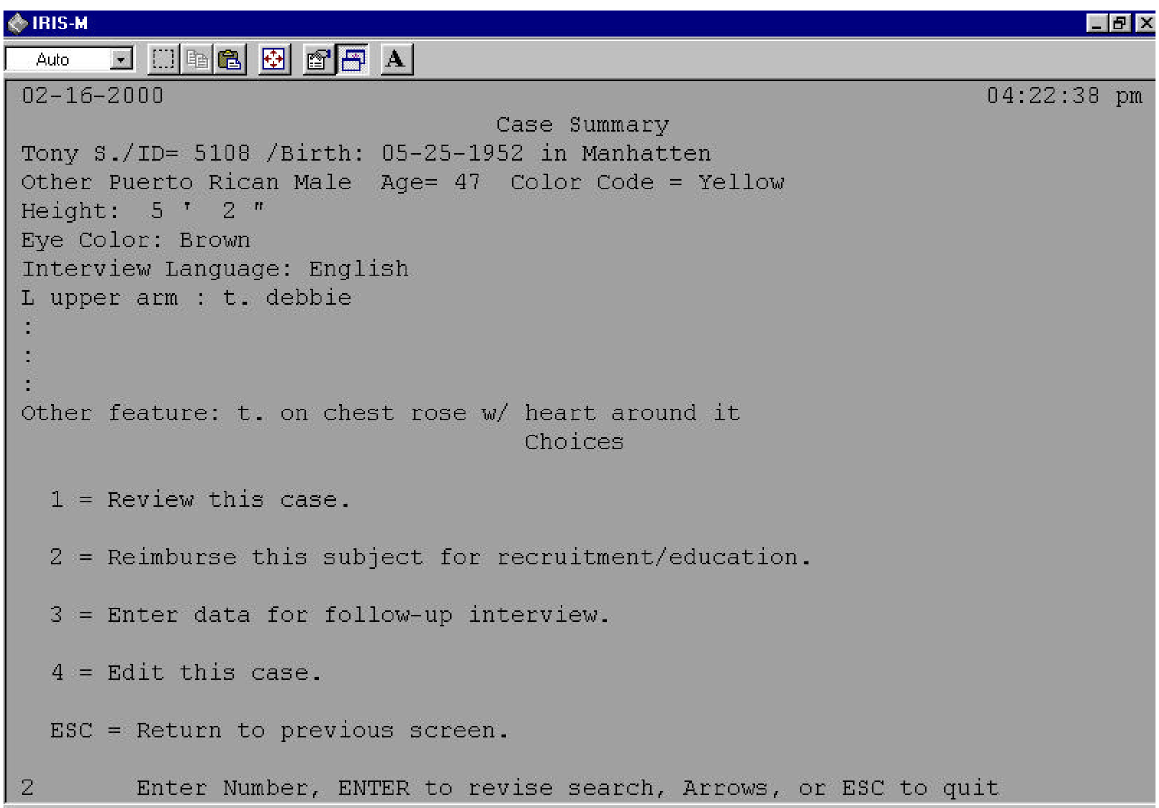
Now the Health Educator searches the Features database for any former respondent with a tattoo of "Debbie," and IRIS finds the following:

Figure 3: IRIS Screen



José is clearly not #1, a female, but he could be #2. She calls up the Case Summary of "Tony S.":

Figure 4: IRIS Screen



In height, José is a match. She also learns that Tony S. has another tattoo ("t."), on his chest. Her check of the visible part of José's chest reveals a tattoo of a rose inside a heart. Thus, it becomes virtually conclusive that José is a former ECHO respondent by the name of Tony S., and a click on "Review this case" reveals that he was last interviewed on September 25, 1995.

The use of IRIS in this case prevented the same subject from cycling through the ECHO Project under a different name, perhaps not just once but several times. Furthermore, the ability to identify people without knowing who they are very likely causes respondents like "Tony S." to spread the word on the street that it is not worth trying to fool the project for a $20 interview reward.

IRIS also proves invaluable in preventing respondent duplication and impersonation when respondents return to the ECHO Project for their follow-up interviews (approximately six months later) and to collect their rewards for educating and recruiting their peers (within one to two weeks after their last interview). Respondents frequently complicate matters by forgetting the pseudonyms they have given to the project, or not remembering whether they have been interviewed before by the ECHO Project. It is therefore crucial for ECHO Project staff to be able to identify each respondent, calculate his or her rewards, and verify the date of the last interview, all without knowing the respondents' real names or social security numbers.

Not all respondents are as easy to identify as Tony S., however; because they may lack such features as Tony's visible tattoos. For example, an IRIS search of the ECHO database of 175 subjects to identify a Puerto Rican male, 5'11," born on December 9, 1971, in San Juan, who has no other discriminating physical features, yields a list of several potential respondents, but no easy or efficient match. *It is in exactly such situations that biometric measurements of respondents' wrists and forearms demonstrate their utility, in relation to other "soft" indices*. After re-measuring this respondent's biometrics, the Health Educator searches on "Features" and retrieves the following match:

Figure 5: IRIS Screen

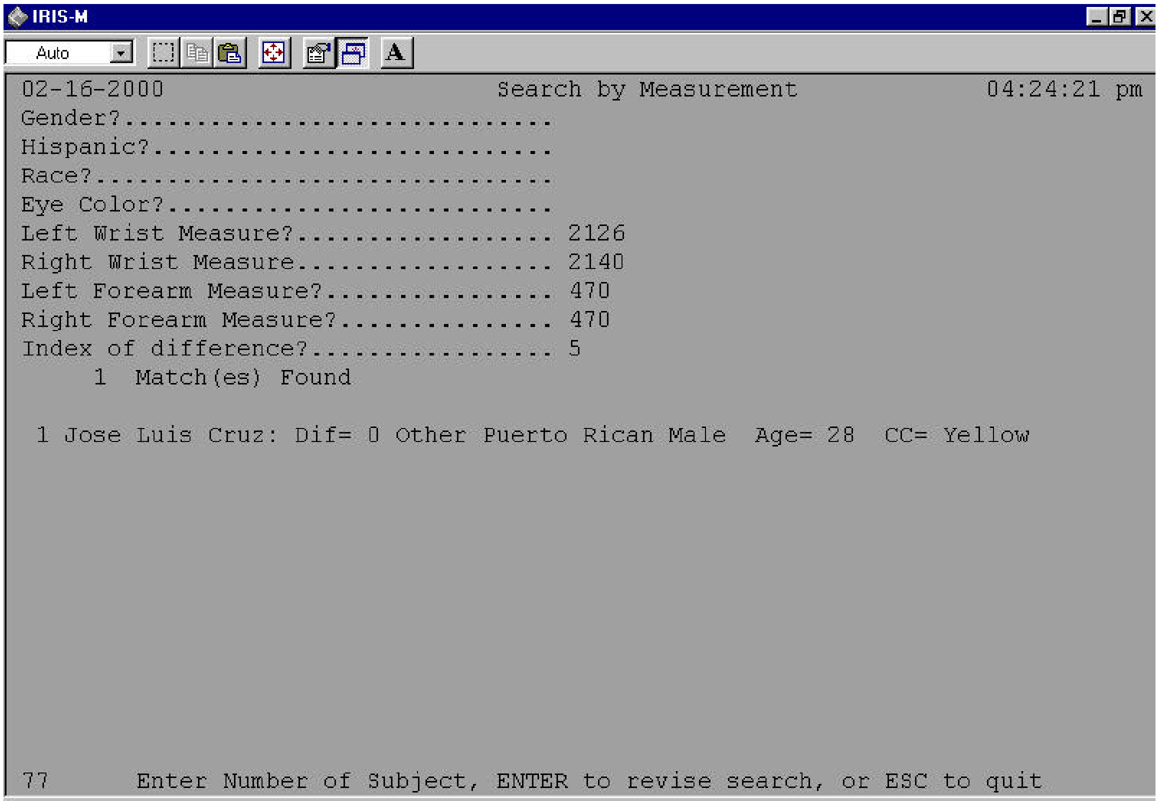
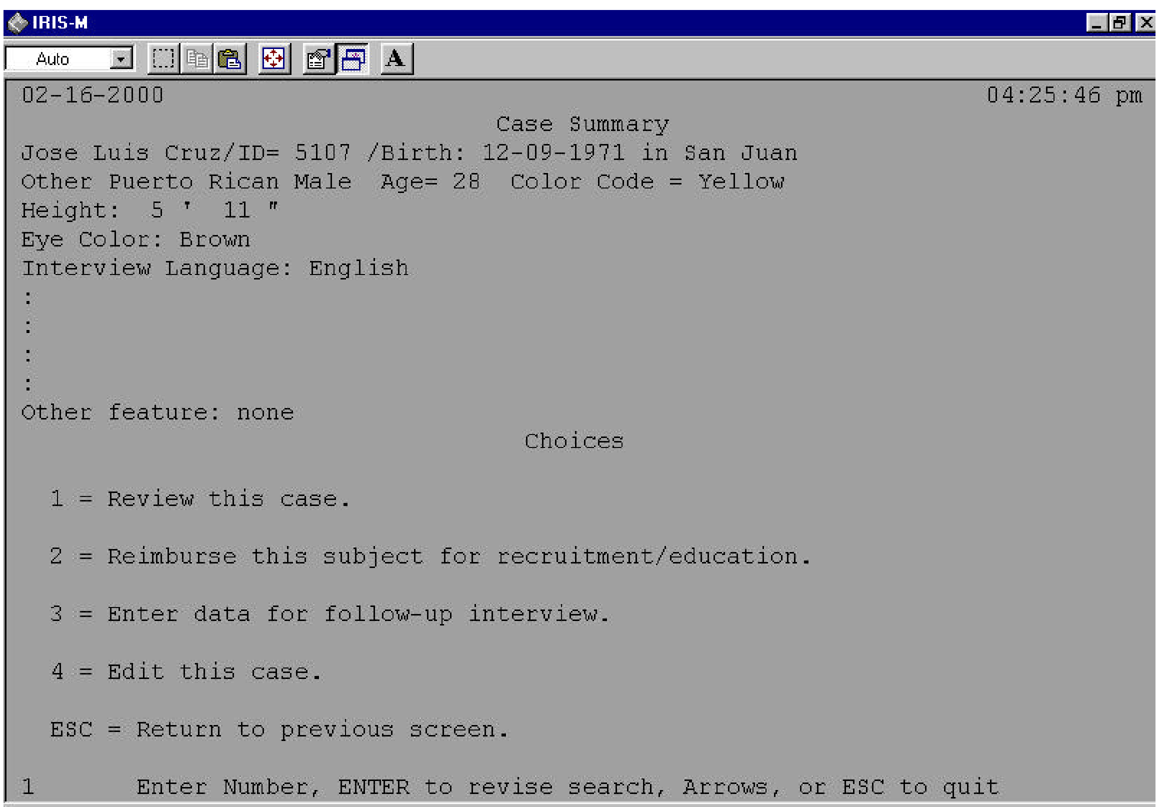


Figure 6: IRIS Screen



Using IRIS, one can enter the data for a specific respondent, and the program produces a list of possible matches, ordered from the most to the least promising. In the case of a perfect match, the target respondent's record appears at the top of the list. In the case of imperfect matches, the target respondent's record appears lower down on the list.

IRIS has been in use since March 1994 as part of the ECHO Project. Measurements are taken at the initial interview and repeated when respondent duplication or impersonation is suspected or when a respondent's identity is unclear. The most common use is to recover pseudonyms that subjects have forgotten. Many of the pseudonyms are whimsical—Freddie Krueger is a common choice—and are particularly easy to forget. IRIS makes it easy to re-identify such respondents quickly.

To date, no respondent has objected to the biometric measurements. Indeed, many respondents welcome their use because it ensures that no one else will receive the rewards they have earned for their recruitment and education efforts in the community: These respondents want assurance that project staff are not mistaking them for some other respondent who may have been unsuccessful in recruiting and educating.

Quantitatively Assessing the Accuracy of the Methodology

Assessing the accuracy of a method for subject identification is difficult when studying a hidden population because, as the Chicago researcher noted above, successful efforts at impersonation or duplication are never detected. Therefore, to assess quantitatively the accuracy of IRIS, we turned to a non-hidden population whose identities were objectively verified, University of Connecticut undergraduates. For the sample of 101 subjects, personal identifiers were drawn first by one staff member and then by another. The information gathered included (1) race/ethnicity, (2) gender, (3) eye color (blue or hazel versus brown), (4) height (plus or minus three inches), and (5) scars, tattoos and other special features. We did not include weight because it is variable over the months that typically elapse between initial and follow-up interviews. Furthermore, HIV-related medical conditions can produce rapid weight change. We also did not include age because it is frequently difficult to judge, for example, whether someone is 28 or 38. Significant weight changes can also alter the appearance of age.

The accuracy of any testing method depends on two factors, sensitivity and specificity. In the case of a test for subject duplication, sensitivity refers to the ability of the test to identify potential cases of duplication. That is, when a candidate for admission to the study is screened, the IRIS program generates a list of potential duplicates. If the candidate was a duplicate, and his or her name appeared on the list, this is a case of a *true positive* (TP). Alternatively, if the duplicate's record was not on the list, the case is a *false negative* (FN). Sensitivity is defined quantitatively as follows:

Sensitivity = TP/(TP+FN)

The test is perfectly sensitive if it always senses the duplicate, that is, if records for duplicate subjects are always placed on the list. In essence, a highly sensitive measure casts a wide net that captures all potential duplicates.

The second determinant of accuracy is the method's *specificity;* that is, are records other than those of the true duplicate placed on the list of potential duplicates? Such subjects are false positives (FP). Alternatively, if a record is not placed on the list, and if that judgment was correct, the case is a true negative (TN). Specificity is defined quantitatively as follows:

Specificity = TN/(TN+FP)

Therefore, the test is perfectly specific if only actual duplicate subjects are placed on the list, and the method thereby avoids false positives. A test is perfectly accurate if it is both perfectly sensitive and perfectly specific, never producing either false negatives or false positives.

When translated into those terms, the principal problem in constructing a method for subject identification is specificity, not sensitivity. Consider, for example, the case of gender. If the candidate is male who is seeking duplicate participation in the study, a list of records of all male subjects will include the subject's record, so the sensitivity will be perfect. However, the specificity would be poor because of the large number of false positives. However, if a second attribute, such as eye color is added, specificity improves because false positives with eye colors different from the candidate are removed. Addition of each new item of information therefore improves specificity.

Table IA: Personal Attributes as Respondent Identifiers		
Personal Attribute	*Specificity*	*Power of Differentiation*
Race	.3393	1.51
Gender	.4939	1.98
Eye Color	.4959	1.98
Height within 3"	.5517	2.23
Scars, Tattoos, Other	.2517	1.34
Combined	.9185	12.27

Table IA reports the mean specificity of each of five directly observable indicators. As thus computed, gender excludes about 50 percent of respondents. This reflects the approximately equal gender breakdown of the sample. Similarly, eye color excludes about 50 percent, reflecting a nearly equal breakdown between respondents with blue or hazel versus brown eyes. By this measure, race is the weakest indicator, excluding on average of only 34 percent of respondents, and height is the strongest, excluding on average 55 percent of respondents. Table IA's third column reports what we term the *power of differentiation*. Where *specificity shortfall* is defined as one less the specificity (i.e., the extent to which specificity is imperfect), the power of differentiation is defined as the reciprocal of this specificity shortfall. As thus defined, power of differentiation is positively related to specificity and hence provides what might appear to be a redundant measure of specificity. However, power of differentiation has an attribute that makes it useful for understanding the effects of combining indicators. When indicators are combined and the indicators are independent of one another, their powers are multiplicative. When indicators are correlated, the combined power is less than multiplicative, and when indicators are perfectly correlated, their combination adds no power.

It is apparent from the data in Table IA that taken individually, none of the five indicators provide an adequate basis for personal identification: their specificities vary from .2517 to only .5517. In combination, however, they do somewhat better. For example, gender and eye color have powers of differentiation of approximately 2 (1.98), and they are virtually unrelated, so their combined power of differentiation is approximately 4 (3.97). Similarly, if all five attributes were unrelated, their combined power of differentiation would be 17.2 (the product of 1.5, 2, 2, 2.2, and 1.3). However, their computed combined power of differentiation is less, only 12.3, because of interdependencies in the attributes: eye color is associated with ethnicity, for example, and height and the presence of scars and tattoos are associated with gender. This reduction in power is analogous to the problem of multicolinearity in regression analysis.

Given that the combined specificity of the five personal attributes is only .9185, it is apparent that except in the smallest of studies, additional information is needed. We surveyed the literature on biometric measurements, seeking measures that were reliable and could be taken without violating people's personal space. Many reliable measures have been identified in the literature. Measures of soft tissue, such as diameter of the waist, tend to be variable; skeletal measures tend to be more reliable. Some of these, such as the width of the head, we rejected as invasive of personal space. Eventually, we settled on two measures, the length of the forearm, measured from elbow to fingertip, and the width of the wrist. The former was measured using a ruler attached to a board with a block against which the elbow could rest when the measurement took place. The latter was measured using digital calipers. Measurements were taken from both wrists and both forearms, for a total of four biometric indicators.

Table IB reports the results of the analysis of the specificity of the biometric indicators. For details, see the Appendix.

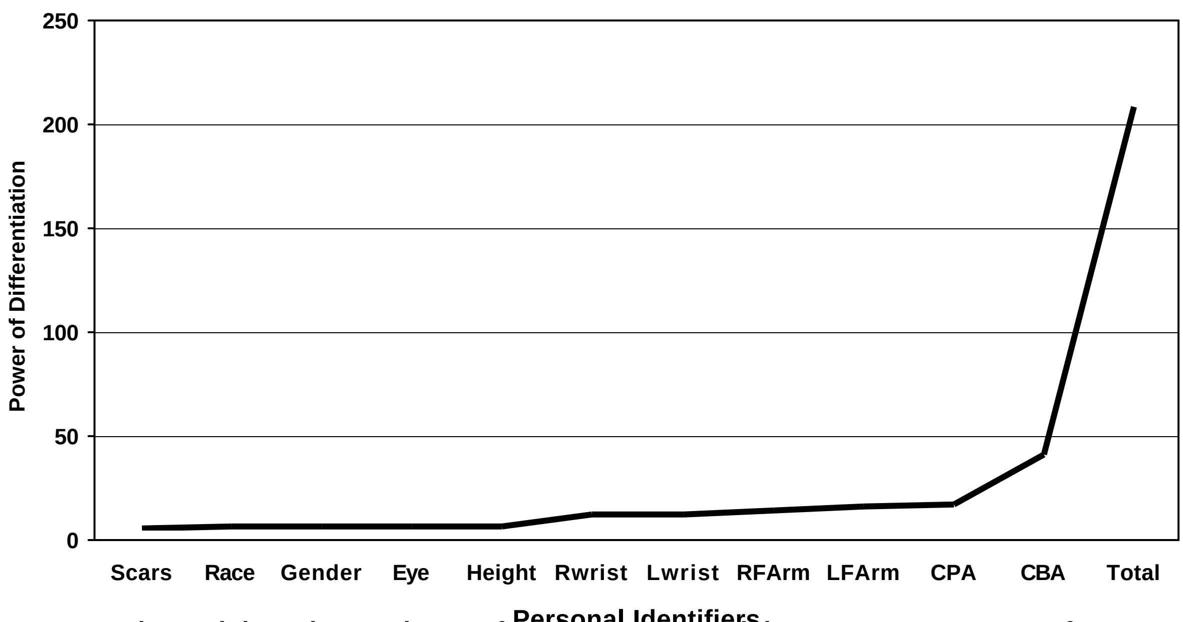
Table IB: Biometric Measures as Respondent Identifiers		
Biometric Measure	*Specificity*	*Power of Differentiation*
Left Wrist	.8718	7.80
Right Wrist	.8651	7.41
Left Forearm	.9110	11.24
Right Forearm	.8973	9.74
Combined	.9730	37.04

In essence, the procedure involved converting the biometric measure to a z-score, and then computing an index of difference defined as the squared difference between the candidate's measure (i.e., a specific subject's follow-up measure) and the measure for each subject in the database (i.e., all subjects' initial measure). Based on this index, each subject in the database was compared for similarity to the candidate. If the candidate's record was the best match, specificity was perfect and there were no false positives. If the candidate's record was not the best match, better matches counted as false positives. Subjects ranked as worse matches than the candidate's record counted as false negatives. Specificity was computed using these counts of false positives and true negatives. Mean specificity was computed by treating each of the 101 subjects as the candidate subject and comparing his or her follow-up measure with all subjects' initial measures. As thus computed, the measurements of the left and right wrists had mean specificity measures of .8718 and .8651, respectively.

When the two wrist measures are combined, their combined specificity increases to .9, which corresponds to a power of differentiation of 10. Note that this is far below the 7.4 * 7.8 = 57.72 that would have resulted had the two wrist measures been unrelated. This occurs because their correlation is high: r =. 961. Nonetheless, combining the two wrist measures does produce an increase in specificity.

The specificity of forearm measurements is somewhat greater, .911 and .8973 for the left and right forearms, respectively. This reflects the greater ratio of bone to soft tissue in the forearm measurement. Yet, like wrist widths, forearm lengths are highly correlated (r = .985), so their combined power is less than multiplicative. However, it is nearly additive because the combined power is 16.1. Finally, when the wrist and forearm measures are combined, the result is a relatively impressive specificity of .973. This corresponds to a power of 37, a figure more than triple the combined power of the personal attributes. This increase in specificity and power reflects the modest correlation (r = .813) between wrist and forearm measures.

Figure 7: Power of Differentiation: Biometrics and Personal Attributes



Rwrist = Right Wrist; Lwrist = Left Wrist; RFArm = Right Forearm; LFArm = Left Forearm; CPA = Combined Personal Attributes; CBA = Combined Biometric Attributes

A comparison of Tables IA and IB shows that the four biometric measures have greater specificity than the five observable attributes: .973 versus .9185. When both types of indicators are combined, power becomes even greater. Were these two types of measures unrelated to one another, the combined power would be 37 * 12 = 444. However, the computed figure is a still impressive power of 204, which corresponds to a specificity of .9951 (see Table IC).

Table IC: Biometric Measures and Personal Attributes as Respondent Identifiers		
Measure	*Specificity*	*Power of Differentiation*
Personal Attributes	.9185	12.27
Biometrics	.9730	37.04
Combined	.9951	204.08

Again, the loss of power results because of associations between the biometric measures and the observational measures (wrist and forearm sizes, for example, being associated with gender and height).

Figure 7 depicts graphically how a set of nine indicators, each of which has low power when considered individually, combine to form an index of far greater power. This combined system of indicators succeeded in identifying respondents 75 percent of the time in the sample of 101 individuals. Even in larger samples, it would reduce the number of false positives to a manageable handful of records that could then be more carefully scrutinized; given a sample size of 1,000, for example, the average number of false positives would be only five. Thus, in combination, the nine indicators provide the basis for a potentially useful method of respondent identification.

The above index can be further refined in either of two ways. First, the accuracy of the indicators could be increased. For example, wrists were measured using digital calipers, and as a result the mean discrepancy between repeated measures was modest, merely .037 inch. In contrast, forearms were measured using a ruler, and the mean discrepancy between repeated measures was more than four times greater, .169 inch. If the forearms could be measured as reliably as the wrists, the increase in power of differentiation would be substantial.

Second, new biometric indicators could be added. For example, height could be measured far more precisely. Recall that height plus or minus three inches had a power of differentiation of 2.2. A three-inch tolerance was chosen because we assumed that this was the range that could be reliably assessed merely by inspection: a respondent who was actually 5’5”, for example, could be reliably placed within a range of 5’2” to 5’8”. Were height measured to an accuracy of one inch, its power of differentiation would increase to 13.1, and the power of the combined measures would increase to an impressive 1,111. The disadvantage is that equipment for measuring height is cumbersome, and use of this measure would therefore be impractical in research involving interviews in the field. Another possibility is measuring the difference between the pupils with a pupilometer (a device that resembles a pair of binoculars).

CONCLUSION

In this paper we focused on a common but neglected methodological problem and source of sampling bias in studies of hidden populations: respondent duplication and impersonation. IRIS, a computer-based information-processing system for recording and retrieving a combination of non-sensitive bio-metric measurements (height, length of forearms and wrist widths), and visible physical characteristics (scars, birthmarks, tattoos, eye color, ethnicity and sex), has been found to reduce the problem substantially, if not completely. Taken together, the accuracy of these multiple indicators for identifying any given respondent is high, without requiring the subject produce positive identification. With the use of a laptop computer, IRIS is also fast and easy to use under field conditions, as in community storefronts where respondents are screened for initial and follow-up interviews, or when they return for subsequent services or to receive respondent fees.

A step-wise photographic tour of the ECHO Project’s peer-driven intervention process is available at the following website, as well as a free copy of the IRIS software for downloading: www.ucc.uconn.edu/~wwwsoci/echo.html.

ENDNOTES

[1] In Figure 2, "**CC =**" refers to a six-month color-code scheme used by the ECHO project to help remind subjects when their next sixth-month follow-up interview was due. Each month the project staff would place a different colored flag in the storefront window; eg., "red" for the months January and July. Thus, subjects first interviewed in January were informed that the next time they saw a red flag in the window, they would know it was time for them to return to the project for their sixth-month follow-up interview, for which they would be paid $30.00.

REFERENCES

Biernacki, P. and D. Waldorf. 1981. Snowball Sampling: Problems and Techniques of Chain Referral Sampling. *Sociological Methods and Research* 10: 141–63.

Broadhead, R.S. and K.J. Fox. 1990. Takin' It to the Streets: AIDS Outreach as Ethnography. *Journal of Contemporary Ethnography* 19 (October):322–48.

Broadhead, R.S. and E. Margolis. 1993. Drug Policy in the Time of AIDS: The Development of Outreach in San Francisco. *Sociological Quarterly* 34 (3):497–522

Broadhead, R.S. and D.D. Heckathorn. 1994. AIDS Prevention Outreach among Injection Drug Users: Agency Problems and New Approaches. *Social Problems* 41:473–95.

Broadhead, R.S., D.D. Heckathorn, J-P.C. Grund, L.S. Stern and D.L. Anthony. 1995. Drug Users versus Outreach Workers in Combating AIDS: Preliminary Results of a Peer-Driven Intervention. *Journal of Drug Issues* 25, #3:531–64.

Broadhead, R.S., D.D. Heckathorn, D.L. Weakliem, D.L. Anthony, H. Madray, R.J. Mills and J.Hughes. 1998. Harnessing Peer Networks as an Instrument for AIDS Prevention: Results from a Peer-Driven Intervention. *Public Health Reports* 113 (June) Supplement 1:42–57.

Cunningham, R.M., L.B. Cottler and W.M. Compton. 1996. Are We Reaching and Enrolling At-Risk Drug Users for Prevention Studies? *Journal of Drug Issues* 23 (3): 541–60.

Frank, O. and T. Snijder. 1994. Estimating the Size of Hidden Populations Using Snowball Sampling. *Journal of Official Statistics* 10:53–67.

Heckathorn, D.D. 1997. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems* 44(2): 174–99.

Heckathorn, D.D., R.S. Broadhead, D.L.Anthony and D.L. Weakliem, D. L. 1999. AIDS and Social Networks: Prevention through Network Mobilization. *Sociological Focus*. 32(2): 159–79.

Hser, Y.-I. and M.D. Anglin, 1993. Prevalence Estimation Techniques for Drug-Using Populations. *Journal of Drug Issues* 23 (2), Special Issue :163–360.

Spreen, M. 1992. Rare Populations, Hidden Populations, and Link-Tracing Designs: What and Why? *Bulletin de Methodologie Sociologique* 6:34–58.

Watters, J.K. and P. Biernacki. 1989. Targeted Sampling: Options for the Study of Hidden Populations." *Social Problems* 36(4):416–30.

Wiebel, W.W. 1988. Combining Ethnographic and Epidemiologic Methods in Targeted AIDS Interventions: The Chicago Model. In R.J. Battjes and R.W. Pickens (Eds.), *Needle Sharing among Intravenous Drug Abuse: National and International Perspectives*. National Institute on Drug Abuse Research Monograph No. 80, pp. 137–50. Washington D.C.: U.S. Government Printing Office.

Appendix

This appendix describes the computation procedures for analyzing the biometric measures. To simplify, we will consider biometric data from only three respondents. Each respondent's biometric measures are defined by a four-touple, i.e., the left wrist measurement, the right wrist measurement, the left forearm measurement, and the right forearm measurement, respectively. The following are the biometric measurements from the initial contact with each of three respondents:

2.47	2.337	19.8	19.6
2.362	2.281	18.4	18.5
1.959	2.077	18.3	18.3

These are then to be compared with a set of biometric measurements taken from the first subject's follow-up measurements.

2.287	2.259	19.9	20

The first step in identifying the subject is to convert the measures to z-scores. Here we use the means and standard deviations from the entire sample. The means and standard deviations are as follows: Left Wrist = 2.04 ±. 2; Right Wrist = 2.07 ± .19; Left Forearm = 18.11 ± 1.38; and Right Forearm = 18.2 ± 1.38. The z-scores are as follows:

2.15	1.4052631579	1.2246376812	1.0144927536
1.61	1.1105263158	0.2101449275	0.2173913043
-0.405	0.0368421053	0.1376811594	0.0724637681

Second, the target measures are converted to z-scores.

1.235	0.9947368421	1.2971014493	1.3043478261

The third step is to compute the index of differences between the candidate and the other measures. This index consists of the squared difference between the z-scores for the candidate and the other measures. That is:

0.837225	0.168531856	0.0052509977	0.084015963
0.140625	0.0134072022	1.1814744802	1.1814744802
2.6896	0.9175623269	1.3442554085	1.5175383323

Note that with respect to the wrist measurements (columns 1 and 2 in the above matrix), the best match to the target is respondent #2, because the index of difference is minimal, e.g., .14 < .84 and 2.7. In contrast, with respect to the forearm measurements (columns 3 and 4), the best match is respondent #1, e.g., .005 < 1.18 and 1.34.

To obtain the index of differences based on all four biometric measures, the next step is to sum across the rows. This yields the combined indices:

1.0950238167

2.5169811625

6.4689560677

As is apparent by inspection, the index is smallest for the first of the three respondents (i.e., 1.1 < 2.5 and 6.5). Despite the conflict between the wrist and forearm measurements, this is the best judgment as to the identity of the target respondent. In essence, forearms prove more decisive than wrists. The respondent was correctly identified, so the number of false positives in this case was zero. The number of true negatives, persons ranking below the respondent, was two, so specificity was 2/(2+0)=1.

It should be noted that this procedure for counting false positives and true negatives uses the candidate's similarity to himself or herself as the breakpoint for differentiating positives from negatives. We did not establish a preset breakpoint because in applications of this method, measurement accuracy can be expected to vary based on the equipment used and the skill of staff. To compute mean specificity, this procedure was repeated for each of the 101 subjects, with each subject's follow-up measures being compared with all subjects' initial measures.