An evaluation of a self-generated identification code

By: Colleen DiIorio, Johanna E. Soet, Deborah Van Marter, Tammy M. Woodring, and William N. Dudley

DiIorio C, Soet JE, Van Marter D, Woodring TM, Dudley WN. An evaluation of a self-generated identification code. Res Nurs Health. 2000 Apr;23(2):167-74. doi: 10.1002/(sici)1098-240x(200004)23:2<167::aid-nur9>3.0.co;2-k. PubMed PMID: 10782875

Abstract:

We describe a self-generated coding form used in a study of HIV prevention practices of college students and provide information on the success rate of matching questionnaires over a 3-year period using the form. The data for this study were from a longitudinal study of HIV risk-reduction practices of college students. In order to match questionnaires over the 3-year study period while maintaining anonymity, participants were asked to complete a self-generated identification form at each data collection point. In the second year of the project, we were able to successfully match 74.3% of the questionnaires to those returned during the first year using 6 to 8 of the code elements on the form, and in the third year, we were able to match 73% of questionnaires to those returned in the second year. Participants for whom questionnaires matched were more likely than participants with unmatched questionnaires to be white students enrolled as underclassmen.

Keywords: self-generated codes | anonymous coding

Article:

Participation in HIV prevention studies often requires respondents to answer questions about their sexual behavior and substance use history. Investigators have shown that participants tend to respond more accurately to sensitive questions if they are not required to disclose their names or other identifying information (Catania, Gibson, Chitwood, & Coates, 1990). Although anonymity is desirable, collecting data from respondents more than once (typically with mailed surveys) and linking participant responses from one administration to the next presents a challenge to maintaining anonymity. A few different approaches have been developed to address

this challenge. Udry and Bearman (1998) describe a method in which they hired a security manager to monitor data collection and to match questionnaires longitudinally. Using this approach, the researchers were able to analyze data from matched questionnaires but were never informed of the identity of the respondents. In another approach, researchers link data by using participant identification numbers such as school or work identification numbers, social security numbers, or driver's license numbers. Although the identity of the participant is never revealed to the researcher, the use of these identifiers makes it relatively easy for the researcher to locate participants' names by gaining access to government or school records. And it is likely that respondents may not trust the researchers' statements about confidentiality and anonymity. Another approach, developed by Carifio and Biron (1978, 1982), is a method in which each participant generates his/her own identification code each time data are collected. This method, referred to as a self-generated identification code, is becoming an increasing popular means for linking participant responses in longitudinal studies.

The self-generated identification coding scheme requires participants to compose their own identification number. This is usually done by asking participants to respond to a series of questions, the responses to which are combined to form a single unique code. To insure uniqueness, reproducibility, and accuracy, even over extended periods of time, the questions selected should have the following characteristics: The answers should remain constant over time (e.g., month of birth, first initial of mother's maiden name); the information is easy for the respondent to remember and is not likely to be forgotten; the number of items is large enough so that a unique code is developed for each person; the information is not so sensitive that respondents choose to skip an item (e.g., birth month as opposed to birth year); the responses can be easily recorded in alphanumeric format; and finally, the items must be of the form that the respondent is confident that it would be difficult for the investigator to locate all the information needed to "break" a person's unique code.

The self-generated identification code, although a unique solution to maintaining anonymity in longitudinal studies, is not without problems. Researchers note that the process of matching responses is less than perfect, and the success of matching is indirectly proportional to the time between administrations (Kearney, Hopkins, Mauss, & Weisheit, 1984). Success of matching is fairly high (>90%) for questionnaires administered within days of each other but lower (78%) for responses obtained a year apart (Kearney et al.). The percent of successful matching is a concern, because data sets in which responses are not matched or are incorrectly matched are subject to sample bias; that is, participants with unmatched questionnaires may differ from those with matched questionnaires on critical areas such as gender, race, and psychosocial variables.

The way in which respondents differ depending on matched identification codes has been the subject of some investigations. Kearney et al. (1984) found that students with matched questionnaires scored higher on knowledge than those with nonmatched questionnaires. There were no other significant differences on 11 other study variables assessed including self-esteem, attitudes about alcohol use, parental control, and peer support. Grube, Morgan, and Kearney (1989), however, found several differences between respondents whose questionnaires matched and those whose did not match. Respondents with unmatched questionnaires were more likely to be male, of lower socioeconomic status, with more spending money, involved in problem

behaviors, and less closely tied to school and religion than participants whose questionnaires matched.

Although researchers have reported the use of self-generated identification codes for linking participants' responses, few give information about their success rates in matching responses. Moreover, the issue of sample bias with respect to reporting sexual behaviors has not been examined. The purpose of this article is to describe the self-generated coding procedures we used in a study of HIV prevention practices of college students; provide information on the success rate of matching; compare responses of participants for whom we could match data and for those whom we could not; and make suggestions on ways to improve matching.

METHOD

Sample

The data for this study were from a 3-year longitudinal study of HIV risk-reduction practices of college students. In the first year of the study, 8,529 students attending one of six colleges or universities in a large southeastern metropolitan area were randomly selected and sent a study questionnaire. Twenty-four percent of students (n = 2,044) returned completed questionnaires. In Year 2 of the study, we sent questionnaires to 1,607 participants who had returned completed surveys in Year 1 and who were currently enrolled in school. Also in Year 2, a second random sample of students (n = 5,893) was selected from the same six colleges and universities forming a second cohort. In Year 2, 3,517 (49%) completed questionnaires were returned, 1,049 from the first cohort and 2,468 from the second cohort of students. In Year 3, we sent questionnaires to participants (n = 2,389) from both cohorts who returned questionnaires in Year 1 or Year 2, and who were currently enrolled in school. Sixty-three percent (n = 1,493) of questionnaires were returned.

Procedures

Prior to data collection, approval was received from institutional review boards at all participating schools. Questionnaire packets that included the study questionnaire, a cover letter containing the elements of informed consent and instructions, and a card or envelope with a study number on it were mailed to students. They were asked to complete the questionnaire and return it along with the card or in the envelope with the study number on it (procedures varied between Year 1 and Years 2 and 3). The intent of the study was to send questionnaires in Years 2 and 3 only to those students who had returned questionnaires the previous year. In order to determine which students had completed the questionnaires, they were asked to return the questionnaire along with the card with the study number on it or in a blank sealed envelope that was then sealed in the envelope with the study number. When received at the study coordinating center, the card or envelope with the study number was *immediately* separated from the questionnaire. The questionnaire itself had no identifying information, thus, once separated from the card or envelope, it was not possible to determine which questionnaire belonged to which participant. The study number was used to determine which students completed a questionnaire, but it could not be used to determine which questionnaire a particular participant had filled out. In Years 2 and 3, registrar's lists were obtained in order to determine if participants who returned

questionnaires the previous year were enrolled in school. Only those participants who were enrolled in school at the time of the second and third data collection periods were sent a questionnaire packet. As part of the questionnaire packet, students received a cover letter providing information about the informed consent process. The cover letter also described in detail the above procedures.

Measures

In order to match questionnaires over the 3-year study period, participants were asked to complete a self-generated identification form at each time (Fig. 1). The form, which was on the first page of the survey each year, was developed by Damrosch (1986). Briefly, the self-identification code was composed of the following elements: initial of mother's first name, initial of father's first name, number of older brothers, number of older sisters, placement of the initial of their first name in the first or second half of the alphabet, month of birth, odd or even birth year, and first initial of middle name. Prior to using the form, slight modifications were made to clarify requests for information. For example, words that might be easily overlooked were capitalized and placed in bold text (OLDER brothers), and participants were reminded to include half-brothers and sisters in their calculations.

FIGURE 1 IS OMITTED FROM THIS FORMATTED DOCUMENT.
Figure 1. Self-generated Coding Form. From "Ensuring Anonymity by Use of Subject-Generated Identification Codes," by S. P. Damrosch, 1986, *Research in Nursing & Health, 9,* pp. 61–63.

The survey questionnaires consisted of variables assessing personal characteristics and psychosocial and behavioral variables to test the study hypotheses. The personal characteristics included in the present analyses were gender, race/ethnicity, and academic status. The primary purpose of the parent study was to assess HIV risk-reduction practices using a social cognitive framework. The survey included measures of sexual behaviors and factors believed to be associated with sexual behaviors. Variables chosen for the present analyses were those that were included in all three surveys. These variables were: condom use self-efficacy; outcome expectancies for safer sex practices; willingness to take risks; initiation of vaginal, oral, and anal sex; discussion about safer sex; condom use behaviors; use of alcohol and drugs; and combining alcohol and/or drugs with sexual intercourse. Condom use self-efficacy was a 12-item scale (Soet, DiIorio, & Dudley, 1998). The outcome expectancies for safer sex scale was composed of three subscales: outcome expectancies for discussion, for abstinence, and for condom use (Soet et al.). Willingness to take risks was measured by a 4-item scale. Initiation of sexual intercourse was measured by asking participants to indicate the age at which they first had vaginal, oral, and anal sex. Participants could also respond "never had" for each of these three items (DiIorio, Dudley, & Soet, 1998). Safer sex discussion was measured by a 7-item subscale of the Safer Sex Behavior Questionnaire (SSBQ), and condom use behaviors were measured by a 5-item subscale of the SSBQ (DiIorio, Parsons, Lehr, Adame, & Carlone, 1992). A3-item scale was used to assess frequency of alcohol and drug use, and a 4-item scale was used to assess combining alcohol and drugs with sex (DiIorio, Dudley, & Soet).

RESULTS

Using questionnaires returned in Year 2 for Cohort 1, questionnaires were matched with those returned in Year 1 using the self-generated identification code. A total of 672 questionnaires returned in Year 2 were matched using all 8 code elements (Table 1). A visual inspection of the code elements of the questionnaires that were not matched revealed that an additional 105 (9.8%) could be matched on 7 elements, and another 20 (1.9%) could be matched on 6 code elements. The questionnaires matched with fewer than 8 elements were required to match on the month of the respondent's birth and on age, gender, race, and school attended. There were 222 (20.7%) questionnaires that could not be matched on 6 or more code elements, and those were classified as unmatched. Finally, there were 53 (4.9%) questionnaires in which the self-identification code form was not completed in Year 2.

Table 1. Number and Percent of Questionnaires Matched for Study Year 2 and Year 3

	Year 2		Year 3			
			Frequency by Total and by Cohort			
Group	Frequency	Percent	Total	Cohort 1	Cohort 2	Percent for Total
Matched on all 8 elements	672	62.7	897	282	615	61.2
Matched on 7 elements	105	9.8	140	32	108	9.5
Matched on 6 elements	20	1.9	32	11	21	2.2
Matched on 5 elements			1	0	1	.07
Unmatched	222	20.7	353	—[a]	—[a]	24.1
No information	53	4.9	43	—[a]	—[a]	2.9
Total	1,072		1,466			

[a]As we did not need to distinguish participants in Cohort 1 from those in Cohort 2 for the parent study, returned questionnaires from members of both cohorts were mixed together. This mixing made it impossible for us to distinguish questionnaires collected from members of Cohort 1 from those of Cohort 2 for this portion of the analysis.

Table 2. Number of Errors Per Code Element on the Self-Generated Identification Forms in Study Year 2 and Year 3

Code Element	Year 2 Frequency ($N = 125$)[a]	Year 3 Frequency ($N = 173$)[a]	Total Frequency ($N = 298$)[a]	Percent of Total
Number of older brothers	28	58	86	24.4
Number of older sisters	27	54	81	23.0
First letter of father's first name	36	36	72	20.5
First letter of mother's first name	27	16	43	12.2
First name in the first or second half of the alphabet	15	23	38	10.8
Year of birth (odd or even)	10	12	22	6.3
Letter of middle initial	2	8	10	2.8
Month of birth[b]	0	0	0	0
Total	145	207	352	

[a]Some surveys had more than 1 error. [b]All identification codes were required to match on this element.

In Year 3, we sent questionnaires to all participants who responded in Years 1 or 2. Because we did not need to distinguish participants in Cohort 1 from those in Cohort 2 for the parent study,

returned questionnaires from members of both cohorts were mixed together. The mixing of questionnaires, however, made it impossible for us to distinguish questionnaires collected from members of Cohort 1 from those of Cohort 2 for the present analysis. We assumed that questionnaires that could be matched for all three years belonged to Cohort 1, and those that could be matched for Years 2 and 3, but not for Year 1 and Year 2, belonged to Cohort 2 respondents. And because questionnaires were mixed together, we were not able to determine whether questionnaires without matches belonged to members of the first or second cohort. This uncertainty is reflected in Table 1 in which the number of unmatched questionnaires received in Year 3 is combined.

In Year 3, we were able to match 897 questionnaires, representing 61.2% of all returned questionnaires in Year 3, on all 8 elements; 140 (9.5%) matched on 7 elements; 32 (2.2%) matched on 6 elements; and 1 (.07%) matched on 5 elements. We were not able to match 353 (24.1%) questionnaires, and for 43 (2.9%) questionnaires the information to develop the code was not completed. For Cohort 1, we matched 282 questionnaires for all three years on all 8 elements, 32 questionnaires on 7 elements, and 11 questionnaires on 6 elements (along with age, race, gender, and school attended). For the second cohort, 615 questionnaires matched on 8 elements, 108 questionnaires matched on 7 elements, 21 questionnaires matched on 6 elements, and 1 questionnaire matched on 5 elements. The overall matching rate for Year 3 was 73%.

In the next set of analyses, we sought to determine if the participants for whom we could match questionnaires were different from those for whom we could not match questionnaires. For the first analysis, the questionnaires obtained for Cohort 1 in Year 2 were divided into four groups: (a) matched on all elements, (b) matched on 6 or 7 elements, (c) unmatched, and (d) no code information provided. Using chi square analysis, the four groups were compared on personal characteristics of gender, race, academic status, and initiation of sexual intercourse. Significant differences were noted on race and academic status. African American respondents and those who classified themselves as "other ethnicity" were overrepresented in the unmatched group $\chi^2(12, N = 1{,}064) = 36.08, p < .001$; graduate students were also overrepresented in the unmatched group, and seniors and graduate students were overrepresented in the group that did not complete the self-identification form, $\chi^2(15, N = 1{,}061) = 39.511, p < .001$.

In the next analysis, the same four groups were compared on the major study variables that were included in the survey for all 3 years. Using analysis of variance, the *F* statistic for comparison of outcome expectancies for discussion for the four groups was significant, $F(3, 1045) = 2.661, p = .047$. However, post hoc comparisons using Sheffé revealed no two groups that were significantly different from each other. Using analysis of variance, a significant difference also was observed among the groups for alcohol and drug use, $F(3, 1043) = 3.62, p = .013$. Post hoc comparison revealed a significant difference between the group of respondents whose questionnaires matched on all elements and the group of respondents whose questionnaires did not match such that those in the matched group had a higher mean alcohol and drug use score than those in the unmatched group. Additional analysis of individual alcohol and drug use items showed that the results were significant only for the two alcohol use items. Because African American participants were less likely than other participants to use alcohol and drugs and were also overrepresented among respondents whose questionnaires did not match, additional analyses

were conducted to determine if there was a matching group by race interaction. Those analyses revealed no interaction effects.

Similar analyses were completed for matching for Year 2 to Year 3. The results showed no differences by gender. However, African American participants were overrepresented among participants whose questionnaires did not match, $\chi^2(15, N = 1{,}131) = 40.770$, $p < .001$, and participants who had never had vaginal intercourse were more likely than those who had to return questionnaires with matching codes, $\chi^2(3, N = 1{,}141) = 16.928$, $p < .001$. With respect to the study variables, participants whose questionnaires matched were more likely to report use of alcohol or drugs than participants whose questionnaires did not match, $F(3, 1129) = 3.177$, $p = .023$.

A third set of analyses was conducted to determine if the reliability coefficients of scales differed according to group classification: (a) matched, (b) matched on 6 or 7 elements, (c) unmatched, and (d) no code information provided. Using Year 2 data, the results of the analyses showed that the reliability coefficients were fairly stable across groups for the nine scales used in this assessment. Overall the coefficients ranged from .66 to .95 with only 2 scales having alpha coefficients below .74. The differences between groups ranged from .03 to .12 with a mean difference of .06. The greatest difference was between the group of respondents whose questionnaires did not match ($\alpha = .66$) and those who provided no code information ($\alpha = .78$) on the risky behaviors scale. The smallest difference among groups was .03 on the outcome expectancy for discussion scale.

The final analysis was to determine which code elements were most prone to discrepancies in matching. In this analysis, we used questionnaires that matched on 6 or 7 elements. In addition to matching on 6 or 7 elements—one of which was birth month—they were also required to match on age, gender, race, and school attended. We found that most matching discrepancies occurred in reporting the number of older brothers and sisters (Table 2). The second major source of matching discrepancies occurred in providing the father's and the mother's initial of their first name. Discrepancies were least likely to occur in the provision of personal information including the placement of the first letter of their first name in the first or second half of the alphabet. The fewest discrepancies occurred with designating the first initial of the middle name.

DISCUSSION

This report is one of only a few reports to contain an evaluation of the use of self-generated identification codes. The results indicate that a substantial proportion of questionnaires can be successfully matched over a 12-month period using the self-generated identification form. However, the results also indicate that it might be unrealistic to expect close to 100% matching over this time period. In Year 2 of the study, we were able to successfully match 62.7% of questionnaires using all 8 elements of the code, and we were able to match an additional 11.7% using 6 or 7 elements of the code in addition to age, race, gender, and school. In Year 3, we were able to match 61.2% of questionnaires, and an additional 11.8% when we matched 5–7 elements. Our percentage of matched questionnaires is similar to that of Kearney et al. (1984), who matched 78.1% of their questionnaires collected one year apart. Kearney et al. also were able to

increase their percent of matches from 45.8% to 78.1% by including questionnaires matched on 6 of their 7 code elements.

In Year 2, we were able to successfully match 74.3% of the questionnaires and in Year 3, 73% of questionnaires. Although about one fourth of participants did not complete the self-generated identification form correctly at least one time, there were few significant differences in the variables that were of primary interest to the study. Thus, participants who completed the self-identification form accurately both times were just as likely as those who did not to express similar levels of condom use self-efficacy, outcome expectancies for safer sex, to discuss safer sex with their partners, and to use condoms. They were also equally likely to be risk-takers and to combine alcohol or drugs with sex.

We did, however, find significant differences in alcohol and drug use, initiation of sexual intercourse, race, and academic status between groups. In both years, participants who provided sufficient information to match their questionnaires were more likely to report more frequent use of alcohol and, in Year 3, to report not having initiated sexual intercourse. Respondents who provided accurate information for both data collection periods were likely to be white and enrolled in undergraduate studies. McAlister and Gordon (1986) also found that minority participants were overrepresented among their respondents for whom questionnaires did not match. Much has been written about the distrust of research among African Americans (Herek & Capitanio, 1994; Thomas & Quinn, 1991). Thomas and Quinn noted that this distrust stems in part from the Tuskegee Syphilis Study, a study in which many African Americans were denied treatment for syphilis so that researchers could study the long-term consequences of the disease. The legacy of the Tuskegee study and lingering doubts about the intentions of researchers might be a factor in the reluctance of African American participants to record data essential for matching their questionnaires. Understanding the concerns of African Americans about participating in research, particularly longitudinal studies, is important in encouraging full participation.

Our results also showed that information requested about the respondent was less prone to error than information requested about family members. Using the questionnaires matched with 6 or 7 elements, we found that most matching discrepancies occurred with the number of older brothers and sisters. Although we highlighted the word “OLDER” and reminded participants to include half brothers and sisters in their count, participants might not have been attentive to these directions. The second major source of discrepancies occurred in providing the father’s and the mother’s initial of their first name. Many discrepancies could be accounted for by a “given” name versus “nickname” factor. That is, one year the respondent might have given the initial of the father’s given name (e.g., William, Robert) and the next year, his nickname (e.g., Bill, Bob). This same pattern was evident for some of the discrepancies noted in the mother’s first initial code element.

Discrepancies were least likely to occur in the provision of personal information, including the placement of the first letter of their first name in the first or second half of the alphabet. Although participants were likely to know whether the first initial of their first name was in the first or second half of the alphabet, inattention to either the instructions or recording the information might have caused them to give information for their last rather than first name.

Inattention to instructions might also have been responsible for discrepancies in birth year being recorded as odd or even. The fewest discrepancies occurred with designating the first initial of their middle name. Because we required that all questionnaires match on birth month, no discrepancies were recorded for this variable. As in our results, Kearney et al. (1984) found that requests for information about number of older siblings and father's first initial were associated with the greatest number of discrepancies in matching.

To improve code elements, a researcher might ask respondents for more personal information such as full date of birth or a portion of their social security number or school ID (e.g., the first three or the last four digits). However, asking for personal information increases the chance that respondents will become more concerned that the investigator could determine their identity. Another approach to improving the Subject-generated Coding Form developed by Damrosch (1986) might be to improve the directions to prevent common matching discrepancies. For example, reminding participants to circle the first letter of a father's given name rather than his nickname might decrease the number of discrepancies related to this code element.

The results indicate that investigators who use an anonymous coding form can expect to retain about 70% of their sample over a 1-year period. Providing detailed instructions for completion and highlighting key points in the directions might reduce careless errors. Efforts to build trust with participants might allay fears and suspicion about misusing the research data. Finally, although sample bias is a major concern in this area of research, we found no differences in social cognitive variables or most demographic characteristics. Contrary to our expectation of more conservative behavior, we found greater alcohol use among participants whose questionnaires matched. However, in the Year 3 analysis, participants whose questionnaires matched were more conservative regarding sexual intercourse.

References

Carifio, J., & Biron, R. (1978). Collecting sensitive data anonymously: The CDRGP technique. Journal of Alcohol & Drug Education, 23, 47–66.

Carifio, J., & Biron, R. (1982). Collecting sensitive data anonymously: Further findings on the CDRGP technique. Journal of Alcohol & Drug Education, 27, 38–70.

Catania, J.A., Gibson, D.R., Chitwood, D.D., & Coates, T.J. (1990). Methodological problems in AIDS behavioral research: Influences on measurement error and participation bias in studies of sexual behavior. Psychological Bulletin, 108, 339–362.

Damrosch, S.P. (1986). Ensuring anonymity by use of subject-generated identification codes. Research in Nursing & Health, 9, 61–63.

DiIorio, C., Dudley, W.N., & Soet, J.E. (1998). Predictors of HIV risk among college students: A CHAID analysis. Journal of Applied Biobehavioral Research, 3, 119–134.

DiIorio, C., Parsons, M., Lehr, S., Adame, D., & Carlone, J. (1992).Measurement of safe sex behavior in adolescents and young adults. Nursing Research, 41, 203–208.

Grube, J. W., Morgan, M., & Kearney, K.A. (1989). Using self-generated identification codes to match questionnaires in panel studies of adolescent substance abuse. Addictive Behaviors, 14, 159–171.

Herek, G.M., & Capitanio, J.P. (1994). Conspiracies, contagion, and compassion: Trust and public reactions to AIDS. AIDS Education & Prevention, 6, 365–375.

Kearney, K.A., Hopkins, R.H., Mauss, A.L., & Weisheit, R.A. (1984). Self-generated identification codes for anonymous collection of longitudinal questionnaire data. Public Opinion Quarterly, 48, 370–378.

McAlister, A., & Gordon, N.P. (1986). Attrition bias in a cohort study of substance abuse onset and prevention. Evaluation Review, 10, 853–859.

Soet, J.E., DiIorio, C., & Dudley, W.N. (1998). Women's self-reported condom use: Intra and interpersonal factors. Women & Health, 27, 19–32.

Thomas, S.B., & Quinn, S.C. (1991). The Tuskegee Syphilis Study, 1932 to 1972: Implications for HIV education and AIDS risk education programs in the Black community. American Journal of Public Health, 81, 1498–1505.

Udry, J.R., & Bearman, P.S. (1998). New methods for new research on adolescent sexual behavior. In R. Jessor (Ed). New perspectives on adolescent risk behavior, (pp. 241–269). Cambridge: Cambridge University Press.