Self-Generated Identification Codes for Anonymous Collection of Longitudinal Questionnaire Data

KATHLEEN A. KEARNEY, RONALD H. HOPKINS, ARMAND L. MAUSS, AND RALPH A. WEISHEIT

RESPONDENT anonymity is often desirable in survey research. There is ample evidence that anonymous respondents tend to give answers that are less self-protective and presumably more accurate than respondents who believe they can be identified (Fidler and Kleinknecht, 1977; Fuller, 1974; Stone, et al., 1977; Tracy and Fox, 1981; Wiseman, et al., 1975–76). Furthermore, practical and legal considerations, such as local research regulations, may also necessitate the use of anonymous questionnaires. When a researcher desires to collect additional data from the same respondents at several points in

Abstract The success of a self-generated identification code for linking longitudinal questionnaire data was examined. The matching procedure developed for linking questionnaires, including a simple technique to compensate for nonidentical codes, yielded a high success rate (92% linkage of cases over a one-month interval and 78% over a one-year interval) and very few incorrectly linked cases. The procedure worked equally well with elementary and high school students, and the resulting samples were representative of the student population on a wide range of measures. Some suggestions are offered regarding the elements comprising self-generated codes.

Kathleen A. Kearney is Research Associate at the Educational Research Centre, St. Patrick's College, Dublin, Ireland; she was a doctoral candidate at Washington State University when this work was completed. Ronald H. Hopkins is Professor and Chair of Psychology at Washington State University. Armand L. Mauss is Professor of Sociology at Washington State University. Ralph A. Weisheit is Assistant Professor of Criminal Justice Sciences, Illinois State University at Normal; he was a postdoctoral research associate in the Social Research Center at Washington State University when this work was completed.

This research was supported in part by Grant No. 5 H84 AA03734 from the National Institute on Alcohol Abuse and Alcoholism to the second and third authors, and was submitted in partial fulfillment of the requirements for the Ph.D. degree in psychology by the first author. Requests for reprints should be addressed to Ronald H. Hopkins, Department of Psychology, Washington State University, Pullman, Washington 99164.

Public Opinion Quarterly Vol. 48:370–378 © 1984 by the Trustees of Columbia University
Published by Elsevier Science Publishing Co., Inc. 0033–362X/84/0048–370/$2.50

time, however, the need for anonymity comes into conflict with the need for data linkability.

Several techniques have been used to enable researchers to connect separate observations belonging to the same respondents without using the respondents' names. One such method uses precoded questionnaires, with respondents told that their names will not be linked to their answers. A second requires that respondents make up some unique code name or number and write it on every questionnaire they complete. Neither of these methods has provided an entirely satisfactory way of collecting longitudinal data anonymously, the first because respondents may not perceive themselves to be anonymous and the second because respondents will almost inevitably have difficulty remembering their ID codes over long intertest intervals.

Recently a more promising technique known as the self-generated ID code has been introduced. Respondents create a code themselves by providing code elements that are well known to them but not to the researchers, such as their own or their parents' initials or birthdates, or specific digits of their street addresses or telephone numbers. Code elements should be chosen which are likely to remain constant for the period of the research and which, in combination, make it unlikely that different respondents will produce duplicate codes. Such a code can, of course, be "broken" under some conditions; for example, a classroom teacher might have access to enough code elements to identify his/her students. Under many research conditions, however, the self-generated ID code is a clever solution to the problem of maintaining linkability while assuring that anonymity cannot be compromised by any practical means.

Relatively little is known about the effectiveness of the self-generated ID code. Only a single methodological study focusing on this innovation has been published to date (Carifio and Biron, 1978). The student respondents in that study indicated that they thought the code was easy to complete; they also perceived themselves to be anonymous, reporting that they would be more likely to provide confidential information if such a code rather than their names were used for identification. With respect to linkage efficiency, Carifio and Biron reported nearly perfect success in matching codes over short intertest intervals of one to three days. Although self-generated codes have been used for identification in a number of studies with more typical intertest intervals of three months to two years (e.g., Andrews and Kandel, 1979; Groves, 1974; Josephson and Rosen, 1978; Kandel, 1973; Kandel, et al., 1978; Stuart, 1974), none of these has provided sufficient information to calculate accurately the matching success rates.

The possibilities of incorrectly linked cases and of sample bias are also important considerations in judging the usefulness of self-generated codes. Although none of the previous studies yields any information about incorrectly linked cases, Kandel, et al. (1978) and Josephson and Rosen (1978) did obtain some evidence of sample bias. They reported that students whose successive ID codes could not be matched were more likely than those who provided matching codes to report marijuana use, school absenteeism, and nonconformist attitudes.

More information is needed regarding the efficiency and effectiveness of self-generated ID codes in longitudinal research. The purpose of our research was systematically to investigate the effectiveness of a particular self-generated ID code used to link student questionnaires collected at intervals of approximately one month and one year. Our procedures permitted us both to determine the efficiency of the code and to investigate sample bias. A simple procedure was also developed for dealing with most of the similar but nonidentical ID codes, and the effectiveness of this procedure was assessed.

Method

SAMPLE AND DATA COLLECTION

The data used in this study were collected as part of an evaluation of a model alcohol education curriculum. These data were taken from questionnaires collected in two urban/suburban school districts during the 1978–79 and 1979–80 school years from students in grades four through twelve.

For some of the junior and senior high students, questionnaires were administered in mass assemblies supervised by teachers and members of the evaluation staff. In most cases, however, the questionnaires were administered in the classes by the students' regular classroom teachers. Teachers were instructed to treat the questionnaires confidentially and anonymously by such means as collecting them in envelopes, etc. Teachers had also been given both verbal and written instructions concerning test administration. These instructions to teachers emphasized particularly the Personal Code section of the questionnaires, instructing the teachers in dealing with special situations (e.g., use "X" if no middle initial, use full names rather than nicknames, etc.).

The data collection procedure also permitted the use of school district ID codes to link the longitudinal data but without additional threat to student anonymity. Self-adhesive labels containing each student's name and school district ID number were distributed to teachers to be used in the testing sessions. Students separated their names from their identification numbers by tearing the labels in half, and then affixed the portion containing the number to the inside front cover of the questionnaire booklet.

Data on the short-term effectiveness of the self-generated ID code were obtained from students who completed pretests and posttests at an interval of three to four weeks. Long-term data came from students who completed the questionnaire once each year. Only those cases which could be linked by school-district ID were used in the analyses described here; there were a total of 130 such cases for the shorter (one-month) intertest interval and 383 cases for the longer (one-year) intertest interval.

INSTRUMENTS

All the questionnaires began with an identical box labeled Personal Code that contained the items used to construct the self-generated code. The code was made up of seven elements: the first letter of the student's middle name, the first letter of the student's mother's and father's first names, the student's sex, birth month, and racial/ethnic category, and the number of older brothers and sisters in the student's family.

The following questionnaire variables were used to assess the bias of longitudinal samples: (1) knowledge, the z-transformation of the number of correct responses to multiple-choice questions about alcohol and alcoholism; (2) self-esteem, measured using a modified version of the Coopersmith Self-Esteem Inventory (Coopersmith, 1968); (3) attitudes, responses to Likert-type items reflecting opinions about alcohol use and abuse; (4) drinking behavior, responses to items asking about experience with alcohol; and (5) marijuana use, response to an item about frequency of marijuana use during the past year (asked only in junior and senior high). The remaining seven questionnaire variables were all based on Likert-type questions, and included (6) parental support, (7) parental control, (8) peer support, (9) peer control, (10) school performance, (11) school satisfaction, and (12) general happiness.

MATCHING PROCEDURE

Preliminary examination of the data indicated that the percentage of exact matches of self-generated ID codes was on the order of 65 percent, an undesirably low figure. Thus, we sought a procedure for linking data when the codes were sufficiently similar that they could reasonably have come from the same student, but were not an exact match.[1] Further examination of the data revealed that acceptance of "off-one" matches (pairs of codes that differed on any *one* of the seven elements, including missing elements) would bring the matching rate up to a quite acceptable level. Linkage by codes different on more than one element was not seriously considered because of the increased possibility for mismatching, i.e., incorrectly linking data. However, a count of the number of matches among the codes produced by *different* students indicated that the probability of a mismatch was extremely low for both the exact matching and off-one matching procedures (see Kearney, 1982, for details). Thus, we adopted the off-one matching procedure and examined its consequences.

Results

EFFICIENCY OF THE SELF-GENERATED ID CODE

The efficiency of the self-generated code was examined by comparing cases linked via this code to those linked by school district ID number, and then to count the numbers of correct matches and mismatches yielded by the self-generated ID code. This analysis was done separately for the short and long intertest intervals. The frequencies and percentages of cases of exact matches, off-one matches, and incorrect matches of the self-generated ID code are summarized in Table 1. The data are presented separately for elementary (grade six and below) and secondary (junior and senior high) students, and for the one-month and one-year intertest intervals.

As can be seen in the lefthand portion of the table, 66.9 percent of the possible cases (at all grade levels combined) were successfully linked after the one-month interval by exact matches of the self-generated ID code. The off-one matching procedure provided a dramatic improvement to 91.5 percent successful linkage, and added only two incorrectly linked cases to the longitudinal sample.

The long-term linkage data for students tested once each year are summarized in the righthand portion of Table 1. It may be seen that,

[1] We explored the possibility of explicitly calculating the probability that the particular combination of code elements shared by two similar codes could occur twice by chance. This would be a relatively simple matter if the distributions of the code elements were mutually independent, since the probability of the combination would be equal to the product of the probabilities of its components, and the obtained percentage frequency of each value of each of the six components could be used as an estimate of its probability. However, analyses showed that several of the code elements were related. In order to accurately estimate the probability of any given self-generated code, it would thus have been necessary to use contingent probabilities, a procedure which was dismissed as being too expensive and complicated.

Table 1. Matching of Self-Generated Codes Over One-Month and One-Year Intervals

	One-Month Linkage			*One-Year Linkage*		
Variable	*Gr 4–6*	*Gr 7–12*	*Total*	*Gr 4–6*[a]	*Gr 7–12*	*Total*
Linked by school-district ID						
Frequency	19	111	130	238	145	383
Linked by self-generated ID						
Exact matches						
Frequency	12	75	87	105	56	161
Percentage success[b]	63.2	67.6	66.9	44.1	38.6	45.8
Off-one matches						
Frequency	6	26	32	84	54	138
Percentage success[b]	31.6	23.4	24.6	35.3	37.2	36.0
Total						
Frequency	18	101	119	189	110	299
Percentage success[b]	94.7	90.9	91.5	79.4	75.8	78.1
Incorrectly linked by self-generated ID						
Frequency	0	2	2	0	2	2
Percentage success[c]	0.0	1.9	1.6	0.0	1.7	0.6

[a] Grade six and below in Year 1.
[b] Ratio of number of cases linked by self-generated code to number of cases linked by district code.
[c] Ratio of number of incorrectly linked cases to number of cases matched on self-generated code.

overall, the off-one matching procedure improved the percentage of linked cases from 45.8 percent (for exact matches) to 78.1 percent, but added only two incorrectly linked cases. The overall success rate of 78.1 percent for long-term matching is significantly less than the corresponding 91.5 percent for short-term matching, $\chi^2(1) = 11.67$, $p < .001$, but is still quite respectable.

SAMPLE BIAS

Sample bias was examined only in the long-term data because of the very small number of unlinkable cases in the short-term data. Students whose questionnaires were and were not linked by the self-generated code were compared with respect to their Year-1 scores on the 12 questionnaire variables described above. The only significant difference was that elementary students whose questionnaires could be linked scored significantly higher on knowledge (mean z-score of 0.12) than those whose questionnaires could not be linked (mean z-score of -0.44), $t(232) = 3.53$, $p < .001$.

ERRORS AND OMISSIONS FOR INDIVIDUAL CODE ELEMENTS

Errors on individual code elements were examined in the long-term data because both the sample size and the error rates were higher in those data than in the data based on a one-month intertest interval.

The percentages of students whose successive self-generated codes differed on each individual component are presented in Table 2. As can be seen in this table, the number of older siblings and the student's racial/ethnic background were the least reliable code elements; the first letter of the student's father's first name was also a relatively unstable component.

Discussion

Exact matching of self-generated ID codes was only partially successful in linking longitudinal data. The off-one matching procedure, however, provided a dramatic improvement resulting in linked samples containing 92 percent of the possible cases over a one-month intertest interval and 78 percent of the possible cases over a one-year interval. Less than 2 percent of the cases in either of the samples were incorrectly linked. The self-generated code seemed to work about equally well with elementary and high school students. Furthermore, students whose questionnaires could be linked were comparable to those whose data could not be linked on a wide variety of variables.

It also seems likely that several improvements could be made in future self-generated codes. For example, detailed examination of the data and follow-up discussions with the teachers suggested that simple modifications to our race item would eliminate two problems, namely, multiple responses and confusion of "Native American" with "born in America." The relatively high error rates in number of older siblings and father's initial (perhaps due to changing family composition), suggest that these elements should be avoided in constructing self-generated codes. The remaining code elements—middle initial, mother's initial, birth month, and sex—had lower error rates and seem to be good choices for use in future codes.

Table 2. Percentages of Errors and Omissions on Each Element of the Self-Generated ID Code

Code Element	*% Missing*[a]	*% Different*[b]	*Total %*
Middle initial	4.4%	6.3%	10.7%
Birth month	2.4	0.0	2.4
Sex	3.4	2.0	5.4
Mother's initial	1.5	6.6	8.0
Father's initial	3.2	13.2	16.3
Older siblings	4.1	18.5	22.7
Race	4.9	15.6	20.5

[a] Percentage of students who left the element blank on one or more of the tests.

[b] Percentage of students who gave different values for the element at different test periods.

The generalizability of our results to different situations and to codes based on different elements will require further test. Nevertheless, linkage success rates in the range obtained here offer considerable promise for researchers conducting longitudinal studies. Further improvement might be achieved in some situations by relaxing the criterion for objective matches (e.g., by permitting "off-two" matches or assuming missing elements to match). Alternatively, more qualitative matching techniques (such as comparisons of handwriting style) might be useful in situations with small sample sizes. We did not consider these options appropriate for our situation because of considerations of objectivity, sample size, and credibility of the assumption that the resulting sample is uncontaminated by substantial numbers of mismatched cases.

It does appear, then, that the self-generated ID code, particularly when used in conjunction with the off-one matching procedure, has the potential for being an excellent solution to the long-standing problem of collecting longitudinal data while protecting respondent anonymity.

References

Andrews, K. H., and D. B. Kandel
1979 "Attitude and behavior: a specification of the contingent consistency hypothesis." American Sociological Review 44:298–310.
Carifio, J., and R. Biron
1978 "Collecting sensitive data anonymously: the CDRGP technique." Journal of Alcohol and Drug Education 23:47–66.
Coopersmith, S.
1968 "The antecedents of self-esteem." San Francisco: Freeman.
Fidler, D. S., and R. E. Kleinknecht
1977 "Randomized response versus direct questioning: two data-collection methods for sensitive information." Psychological Bulletin 84:1045–49.
Fuller, C.
1974 "Effect of anonymity on return rate and response bias in a mail survey." Journal of Applied Psychology 59:292–96.
Groves, W. E.
1974 "Patterns of college student drug use and lifestyles." In E. Josephson and E. E. Carroll (eds.), Drug Use: Epidemiological and Sociological Approaches. Washington: Hemisphere.
Josephson, E., and M. A. Rosen
1978 "Panel loss in a high school drug study." In D. B. Kandel (ed.), Longitudinal Research in Drug Use: Empirical Findings and Methodological Issues. Washington, D.C.: Hemisphere.
Kandel, D. B.
1973 "Adolescent marijuana use: role of parents and peers." Science 181:1067–70.
Kandel, D. B., R. C. Kessler, and R. Z. Margulies
1978 "Antecedents of adolescent initiation into stages of drug use: a developmental analysis." In D. B. Kandel (ed.), Longitudinal Research on Drug Use: Empirical Findings and Methodological Issues." Washington, D.C.: Hemisphere.
Kearney, K. A.
1982 "Collecting longitudinal data anonymously: compensating for respondent

errors in self-generated identification codes." Unpublished doctoral dissertation, Washington State University.

Stone, E. F., M. D. Spool, and S. Rabinowitz
1977 "Effects of anonymity and retaliatory potential on student evaluations of faculty performance." Research in Higher Education 6:313–25.

Stuart, R. B.
1974 "Teaching facts about drugs: pushing or preventing?" Journal of Educational Psychology 66:189–201.

Tracy, P. E., and J. A. Fox
1981 "The validity of randomized response for sensitive measurements." American Sociological Review 46:187–200.

Wiseman, F., M. Moriarity, and M. Schafer
1975–76 "Estimating public opinion with the randomized response model." Public Opinion Quarterly 39:507–13.