Journal of Professional Nursing 35 (2019) 120–123

Contents lists available at ScienceDirect

Journal of Professional Nursing

journal homepage: www.elsevier.com/locate/jpnu

Protecting Student Anonymity in Research Using a Subject-Generated Identification Code[☆]

Megan Lippe[a,*], Bailey Johnson[a], Patricia Carter[b]

[a] University of Alabama Capstone College of Nursing, 650 University Blvd. East, Tuscaloosa, AL 35401, United States
[b] University of Texas at Austin School of Nursing, 1710 Red River St., Austin, TX 78701, United States

ARTICLE INFO

Keywords:
Nursing education research
Education, nursing
Research subjects
Research personnel
Students

ABSTRACT

Background: Within nursing education research, protection of students as human subjects must be the highest priority. This protection can be provided via student anonymity. A subject-generated identification code, comprised of responses to a series of questions, can link data across time points while protecting student anonymity.
Method: Two studies, focused on palliative care education, used a subject-generated identification code to link student data across multiple time points. Refinements to the code were made between studies to further enhance anonymity and response consistency.
Results: The subject-generated identification code fostered linking of student responses across three time points in study one and two time points in study two.
Conclusion: There are many benefits to utilizing a subject-generated identification code in nursing education studies. Researchers must consider the need for a data management expert and balancing transposition errors and the power to differentiate between responses.

Introduction

The protection of human subjects must be a primary concern for every investigator, including those conducting research within academic settings. In order to provide high quality nursing education and motivate students, educators must engage in positive and constructive relationships with students while carefully managing power dynamics (Chan, Tong, & Henderson, 2017a, 2017b). However, management of power differentials must be treated differently when faculty members attempt to recruit their students for research. Comer (2009) cautions against recruitment of faculty member's own students into research due to potential threats to justice, confidentiality, and informed consent. However, should the study require the incorporation of one's students, faculty members must engage in practices that prevent identification of the student by any possible means. For example, enrollment of one's own students in qualitative studies in which the faculty member conducts focus groups or interviews poses a great risk to confidentiality (Comer, 2009). For quantitative data, faculty members should provide students opportunities to decline participation free of retribution, engage in data collection separate from the classroom setting, and ensure students are "given meaningful informed consent" (Comer, 2009, p. 104). Ensuring true confidentiality and anonymity of participants may be one means of reducing students' fear of retribution.

In an effort to protect student confidentiality and anonymity, careful steps must be taken by researchers so that no information is collected that could potentially identify a student. For example, asking for age may lead to the identification of outliers who may be much older or younger than the rest of the sample. Similarly, in mostly homogeneous groups of nursing students, asking for a participant to identify gender might facilitate identification of the few males in the group. However, another challenge facing researchers conducting repeated measures studies is determining the best methods to link data across time points without using identifiable information.

Carifio and Biron (1978) proposed the use of a subject-generated identification code (SGID) to link data across time points while protecting participant anonymity. Participants answer the same series of questions at each time point and the responses are combined into a unique SGID. The SGID has been successfully utilized to link data across multiple time points in nursing research (Damrosch, 1986; Schnell, Bachteler, & Reiher, 2010; Yurek, Vasey, & Sullivan Havens, 2008). The number of questions utilized to make the SGID varied from four (Yurek et al., 2008) to eight (Damrosch, 1986). Schnell et al. (2010) conducted

[☆] This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[*] Corresponding author.
E-mail addresses: melippe@ua.edu (M. Lippe), bajohnson6@crimson.ua.edu (B. Johnson), pcarter@mail.nur.utexas.edu (P. Carter).

https://doi.org/10.1016/j.profnurs.2018.09.006
Received 29 March 2018; Received in revised form 14 September 2018; Accepted 21 September 2018

a series of analyses to determine the optimum number of questions to ensure sufficient variability in SGID. They found that SGIDs with longer codes generated from more items (e.g. 10, 12, and 13-items) had the best ability to distinguish between participants, yielded a high rate of linked responses across time points, but added to the participant burden.

In conducting palliative care educational research with nursing students, we utilized the SGID, adapted from the work of Damrosch (1986), in two different research studies. This paper will present the utilization of SGID in nursing education research, including an exploration of key considerations for educators when utilizing the code in their own studies.

Methods

Study one

The first study was performed at a large, research-intensive, public university in the southwest United States. In a cross-sectional, descriptive study, researchers attempted to assess trends in students' knowledge, perceived competence, and attitudes toward caring for dying patients across one baccalaureate nursing program (Lippe, Jones, Becker, & Carter, 2017). Human subject protection approval was obtained by the institutional review board. All students enrolled in any required nursing course during the fall semester were recruited to participate in the study. The primary investigator for the study taught in baccalaureate nursing courses at this university; therefore, special protections were needed to ensure the anonymity of student participants. Damrosch's (1986) eight-item SGID was utilized at each of three data collection time points. This SGID was based on age on a specific date, first letter father's first name, first letter mother's first name, number of older brothers, number of older sisters, in which half of the alphabet is the first letter of first name (A–M = "first"; N–Z = "second"), month born, and first letter of middle name (Table 1). Although other question options could have been selected, this was our first time utilizing the SGID, so we used Damrosch's eight items without modification. We needed to pilot test the SGID to identify any weaknesses or procedural concerns with the existing SGID.

The eight items were presented at the very beginning of the larger survey at each time point. Both written and electronic (Qualtrics) versions of the survey were available to students. Students were informed that the questions were designed to help link their pre-test and post-test responses. Students were instructed to answer each question, and then enter all responses from the eight items questions into one text box to generate their SGID. A sample of the survey questions are provided in Appendix A.

Study two

The second study was performed at a different large, public university in the southern United States. We used the SGID in a repeated-measures, quasi-experimental study assessing the impact of two separate palliative care simulations on students' perceived competence. Simulations were implemented during the fourth and fifth semesters of the college's professional nursing coursework. The fourth semester simulation focused on the care of a dying patient receiving at-home hospice services. The fifth semester simulation focused on the withdrawal of care of a critically ill patient in an intensive care unit. One of the investigators was the simulation director and had contact with all students, so anonymity was critical. Human subject protection approval was obtained by the institutional review board. The SGID was used to link student responses across four time points (pre- and post-test semester 4, pre- and post-test semester 5).

Table 1
Subject-generated identification code question options.

Age in years on (specific date)?
First letter of your middle name (if none write N)
First letter of your mother's first name
First letter of your father's first name
First letter of grandmother's first name
First letter of own surname
Number of *older brothers* (living and deceased)
Number of *older sisters* (living and deceased)
Number of mother's siblings
First name begins with a letter in the first half of the alphabet (A–M) or the second half of the alphabet (N–Z)? If A–M, write "first"; If N–Z write "second."
Name of the month in which you were born
First letter of own birthplace

We identified a few necessary modifications to the SGID from our first study. A modification was made to protect against common errors identified from study one. Specifically, students were asked to indicate the name of their month of birth to avoid previously identified discrepancies. For example, students were specifically instructed to write "June" if they were born in that month. Having students write the month name, rather than the number, was intended to help enhance the accuracy of the SGID. Students could inadvertently report the wrong corresponding number for their birth month, but would be unlikely to report the wrong name. Any spelling errors could easily be identified and corrected when verifying SGIDs during data cleaning.

Furthermore, the item asking about age was removed to enhance confidentiality of student responses. Several students eligible for participation in this study were outliers in terms of age; inclusion of the age question would allow students to be identified from the data. As such, a seven-item SGID was used for the second study. We did not observe any issues with responding to or transposing the other six items; therefore, these items stayed the same for the second study. The seven-item SGID was placed at the beginning of the electronic (Qualtrics) survey at all time points for both simulations.

Results

Study one

Student responses for all three surveys were loaded into Excel. Written responses were manually input into the Excel workbook and confirmed by a second investigator. Perfectly matching SGIDs were identified for a few surveys across time points (n = 37 between survey 1 and survey 2, n = 24 between survey 1 and survey 3, and n = 17 between survey 2 and survey 3). However, many errors in transposition of the SGID occurred from the individual questions to the final code. For example, a student may have indicated the month of birth was "May" and "5" in separate SGIDs. To identify and correct these errors, an Excel data management expert wrote special coding to facilitate item by item comparisons. Following corrections, the final linked sample across all time points was 26. However, this low sample size is reflective of high sample attrition (Lippe et al., 2017) and not a limitation of the SGID.

Study two

Similar processes for identification of errors were utilized in the second study. Surveys were distributed in fall of 2016 and spring of 2017. For fourth semester students in the fall 2016, there were 18 completed surveys at time one and 16 at time two, with 7 total linked matches. For the fifth semester students in the fall 2016, there were 35 completed surveys at time one and 53 at time two, with 22 total linked matches. During the spring 2017, there were 73 completed surveys at time one and 70 completed surveys at time two, with 56 total linked matches. The unlinked responses reflect students who only completed surveys at one time point. The unlinked responses were able to be analyzed for descriptive statistics, but were not included in repeat measures analyses.

Discussion

The SGID was utilized to link student data in two separate nursing education studies conducted at two different institutions. None of the samples had 100% matching of responses across time points, primarily due to different sample sizes at each time and high attrition for the first study. Other studies in which the effectiveness of the SGID was tested also never achieved 100% linking of responses (Damrosch, 1986; Schnell et al., 2010; Yurek et al., 2008). Therefore, inabilities to link responses primarily reflect issues with participation recruitment and retention, as opposed to accuracy of the SGID itself. The removal of the age item for study two did not adversely impact the efficacy of the SGID in linking responses across time.

There are many benefits to using the SGID in education research. One primary benefit in the utilization of the SGID was the ability to link student responses across time points, while protecting their anonymity. Educators can include their own students in research samples without risk of undue power influences because the student identities are protected. Furthermore, the use of standardized questions prevented students from having to memorize a code or share protected information, such as their campus identification codes. Frequently in repeated measures research, participants are asked to create their own code or are assigned a code, which they must memorize. Investigators then keep a book of participants and their codes, which is useful should the participant be unable to recall their assigned code. The SGID precluded the need for a code book, and allowed participants to remain truly anonymous for all aspects of each study.

Prior to the initiation of the second study, the university's institutional review board implemented a policy change such that researchers could not have students use their campus identification codes in research. Many investigators across campus were required to change their procedures for data collection and analysis to ensure responses could be linked across time points. Our SGID was adopted by many faculty researchers as it had already garnered approval from the institutional review board. In a climate of increasing mandates for research protection of subjects, the SGID serves as an approved method for linking data across time points, especially compared to other traditional methods of participant identification.

Finally, incorporating the SGID at the beginning of each survey added on average 3 minutes to survey completion since the questions ask information that students can easily and quickly recall. Completion times in general were the same for the seven and eight item SGID codes. As a result, there is minimal concern about added burden with the use of the SGID in balance with the added benefit obtained.

Several considerations are needed for researchers who plan to incorporate the SGID in their own studies. First, there is a need for a data management expert to assist with linking responses for large data sets. Each SGID represents a series of numbers, letters, and words. Evaluating accuracy of the transposition of each SGID within large datasets without a coding expert would be extremely time consuming and subject to human error. The use of advanced coding to verify the accuracy of the transposed SGIDs, and identify possible transposition errors, allowed us to maximize the number of linked responses for both studies without extra time requirements. A coding expert will be a valuable tool in maximizing the number of matches made, particularly by identifying transposition errors.

Upon initial evaluation, we were concerned that the SGID may have resulted in missed links in data across time points, particularly since our linked sample sizes were much smaller than the response rates of each time point. For both studies, we carefully analyzed each unlinked SGID to see if some unanticipated issues with transposition or SGID completion prevented our coding from identifying proper links. However, for all unlinked responses, no SGID was found to be remotely similar. For example, there were at least three differences in item responses between SGIDs, if not more. Therefore, we felt confident that the SGID successfully linked responses for all students who actually participated in all study time points. The small sample sizes of linked responses appeared to derive more from recruitment and retention issues. Therefore, we feel the SGID is as useful as other methods for linking data across time points, but is also superior in its ability to protect subject anonymity.

Researchers should carefully consider their populations to determine if age should or should not be included as part of the SGID. Although the removal of the age item shortens the SGID, it should not substantially limit the ability to link data sets in moderately sized samples. When considering the number of items to include to obtain the SGID, Schnell et al. (2010) explained that a balance must be struck between transposition errors and the power to differentiate between responses. For the purposes of the two studies reported, the use of eight and seven item SGIDs seemed to strike an adequate balance.

While we found the SGID to be a useful and effective means of linking responses across time points, we acknowledge that other methods can be utilized. One primary limitation of our studies was that we did not collect information about participants' opinions of the SGIDs. Therefore, our findings stem solely from faculty and researcher perspectives. We also did not compare methods of participant identification in a formal analysis; consequently, our findings are based solely on our experiences with using other methods of participants identification compared to the SGID. Future research is needed to fully explore students' perspectives of the SGID and formally compare participant identification methods.

Conclusion

Educational researchers must carefully consider the data collection methods they utilize to minimize the influence of power on students' decisions to participate. The effective use of the SGID in two separate studies at two universities provides support for the utility of using the standardized questions to allow for linking of anonymous responses in longitudinal or repeat-measures studies.

Conflicts of interest

None.

Appendix A. Subject-generated identification code example

To protect your privacy, these surveys will be collected anonymously. In order to link your responses while maintaining your anonymity, please complete the following questions to generate a unique identification code. You will be asked these same questions for every survey, so you do not need to remember this code. Please CAREFULLY answer the following questions.

What was your age in years on September 1, 2017? ________.

What is the first letter of your mother's first name? ________.

What is the first letter of your father's first name?_______.

How many *older brothers* do you have?___________.

How many *older sisters* do you have?____________.

Does *your* own *first* name begin with a letter in the first half of the alphabet (A-M) or the second half of the alphabet (N-Z)? If A-M, write first; If N-Z write second.____.

What is the month in which you were born?________.

What is the first letter of your middle name? If you have no middle initial, write N._____.

Enter all answers from the above questions to generate your unique identification code.____.

References

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

Chan, Z., Tong, C. W., & Henderson, S. (2017a). Power dynamics in the student-teacher relationship in clinical settings. *Nurse Education Today, 49*, 174–179. https://doi.org/10.1016/j.nedt.2016.11.026.

Chan, Z., Tong, C. W., & Henderson, S. (2017b). Uncovering nursing students' views of their relationship with educators in a university context: A descriptive qualitative study. *Nurse Education Today, 49*, 110–114. https://doi.org/10.1016/j.nedt.2016.11.020.

Comer, S. K. (2009). The ethics of conducting educational research on your own students. *Journal of Nursing Law, 13*(4), 100–105. https://doi.org/10.1891/1073-7472.13.4.100.

Damrosch, S. P. (1986). Ensuring anonymity by use of subject-generated identification codes. *Research in Nursing & Health, 9*(1), 61–63. https://doi.org/10.1002/nur.4770090110.

Lippe, M., Jones, T., Becker, H., & Carter, P. (2017). Student preparation to care for dying patients: Assessing outcomes across a curriculum. *The Journal of Nursing Education, 56*(10), 633–637. https://doi.org/10.3928/01484834-20170918-10.

Schnell, R., Bachteler, T., & Reiher, J. (2010). Improving the use of self-generated identification codes. *Evaluation Review, 34*(5), 391–418.

Yurek, L. A., Vasey, J., & Sullivan Havens, D. (2008). The use of self-generated identification codes in longitudinal research. *Evaluation Review, 32*(5), 435–452.