Developing and Validating a Novel Anonymous Method for Matching Longitudinal School-Based Data

Jon Agley1, David Tidd1, Mikyoung Jun1, Lori Eldridge1, Yunyu Xiao2, Steve Sussman3, Wasantha Jayawardene1, Daniel Agley1, Ruth Gassman1 and Stephanie L. Dickinson4

Abstract
Prospective longitudinal data collection is an important way for researchers and evaluators to assess change. In school-based settings, for low-risk and/or likely-beneficial interventions or surveys, data quality and ethical standards are both arguably stronger when using a waiver of parental consent—but doing so often requires the use of anonymous data collection methods. The standard solution to this problem has been the use of a self-generated identification code. However, such codes often incorporate personalized elements (e.g., birth month, middle initial) that, even when meeting the technical standard for anonymity, may raise concerns among both youth participants and their parents, potentially altering willingness to participate, response quality, or generating outrage. There may be value, therefore, in developing a self-generated identification code and matching approach that not only is technically anonymous but also appears anonymous to a research-naive individual. This article provides a proof...
of concept for a novel matching approach for school-based longitudinal data collection that potentially accomplishes this goal.

**Keywords**
longitudinal data, matching, SGIC, self-generated identification code, anonymous, methodology

**Introduction**
Prospctive longitudinal data collection is an important scientific tool for assessing change; it is ubiquitous across academic fields such as medicine and education, and is part of numerous research and evaluation designs, such as randomized controlled trials and pre-/posttest evaluation studies (e.g., Caruana et al., 2015; Seifert et al., 2010). When analyzing longitudinal data, it is important to be able to match responses across data collection points, as each instance of a failed match contributes to attrition rate, and there are meaningful reductions in statistical power when treating repeated measurement of the same subjects as data from independent samples (e.g., Fradette et al., 2003). However, many longitudinal studies are designed to be anonymous (Audette et al., 2020), increasing the difficulty of successfully matching data from the same participant between data collection points. For the purposes of this study, we conceptualized anonymous data as being data that cannot be traced back to an individual participant (Coffelt, 2017), though this standard can be interpreted differentially.

There are a variety of reasons why it might be important to collect study data anonymously. Participants who believe that their responses will be anonymous may more readily endorse stigmatized beliefs (Fear et al., 2012) and report socially undesirable attitudes or behaviors (Durant et al., 2002; Johnson, 2014; Malvin & Moskowitz, 1983). There is some evidence that confidential and anonymous cross-sectional surveys produce similar data, with the caveat that student trust in the sponsoring and administering institution is likely a meaningful moderator (O’Malley et al., 2000). In a school-based environment, anonymous data collection can be an ethical requirement for a waiver of parental consent from an institutional review board when research is low-risk with potential benefits to participants. This waiver can be important, because requiring active parental consent may reduce sample size and introduce bias at the point of recruitment (Anderman et al., 1995; Dent et al., 1993; Shaw et al., 2014), and disadvantage or discourage participation from high-risk populations (Hetzel et al., 2019; Spence et al., 2015; White et al., 2004).

Though there are several different ways to achieve anonymous longitudinal data collection, most techniques require assignment of a unique identifier, or code, to each participant; in a recent, robust review of methodological issues for coding anonymous participants in longitudinal studies, Audette et al. (2020) determined that self-generated identification codes (SGICs) were likely the most effective means of doing so. SGICs have been used for matching anonymous responses for decades with
varying levels of success (Yurek et al., 2008). However, most of the commonly used elements in SGICs draw personalized components from the respondents’ birthday, own name, and mother and/or father’s name (Audette et al., 2020). While Audette et al.’s (2020) review identified these methods as anonymous, the mere fact that they reference personal information—even though only limited elements are being used—may raise privacy concerns in a school-based data collection setting.

At least one study has raised the possibility that some students will assume that research projects are being sponsored and conducted by the school and that the school may have an interest in using data punitively, producing heightened apprehension about anonymity and concomitant reluctance to participate (Gassman et al., 2016). Furthermore, even though there appears to be a consensus that using individual letters or numbers from personal elements makes reidentification of participants highly unlikely (Audette et al., 2020), parents or students may react to the mere mention of middle initials or birthdays on a survey that is labeled anonymous and assume a breach of trust, or even malfeasance (Gassman et al., 2019). The consequences of such assumptions for public health initiatives and projects can be striking. Controversy over study content or privacy may result in an initiative—including large, federally funded public health surveys—being postponed or canceled (Colias, 2014; Kelley, 2015) or being subject to localized outrage (NBCWashington, 2014). Similar incidents also have occurred with surveys using preassigned student identification numbers (Helms, 2020). In addition to effects on the projects themselves, the emotional toll on public health practitioners and researchers at the center of such controversy is often severe and can result in self-censorship and abandonment of research agendas (Kemper, 2008) and may even manifest as direct harassment and threats to the researcher or practitioner (Hoepner, 2017).

Of course, simply asserting—and ensuring—anonymity is one possible response, but doing so does not address the issue of public perception and is unlikely to inoculate any given project from the risk of outrage or concern (Hoepner, 2017). Thus, in the complex landscape of school-based research, individuals working with schools may wish to match data while ensuring both the reality and appearance of extreme difficulty in data reidentification. This may include, for example, providing a letter to parents outlining the questions that will be asked, including SGIC elements, and why they are being asked in that manner (Agley et al., 2020). However, there is a notable absence of a standard or recommended set of SGIC data elements that can accomplish this purpose. Identification and validation of such a set would strengthen the research community’s ability to collect longitudinal data from youth and adolescents enrolled in school.

**Objective**

This study assessed the viability of a novel matching procedure for a school-based research project using an SGIC that did not use any elements derived from personalized public information (e.g., letters from names, birthdays, addresses), rendering the codes incredibly difficult, if not impossible, to reidentify, and with the intention of providing
added reassurance to participants and observers that anonymity was being protected. The purpose was not to develop an SGIC that performed better than extant SGICs, as several proposed approaches (described later in this article) have achieved excellent matching success exceeding 90%. However, those approaches use multiple personalized elements, and so are unsuitable for researchers who are concerned about protecting themselves from misperceptions or outrage about study anonymity. Thus, the goal of this study was to explore and provide evidentiary proof of concept that it is possible to match school-based longitudinal data in a way that will both preserve privacy and is more likely to be perceived as preserving privacy by excluding personalized data elements.

**Method**

This study was a secondary analysis of matching data generated by the ACT OUT! cluster randomized trial (Agley et al., 2020) and fell under the umbrella of that study’s institutional review board approval.

**Survey Administration**

All matching was accomplished using data collected by surveys within the ACT OUT! study, which focused on fourth-, seventh-, and 10th-grade students. At both baseline pretest and 2-week posttest, all participating classrooms ($n = 76$ classrooms across $n = 12$ schools), regardless of the arm to which they were assigned, were provided with a study packet containing surveys, customized Scantron response forms (with classroom ID, trial arm, and grade level documented on the reverse), information on the study structure (pretest, posttest, and follow-up), a manila envelope, a white envelope, and an administrator checklist. Each study packet also contained a label with duplicative information about the classroom to support cross-validation of forms. To comply with ethics requirements, classroom teachers administered the surveys instead of members of the research team. To accurately compute classroom denominators, teachers were instructed to place all response forms that were handed out—including those returned blank—into the manila envelope and to put unused forms in the white envelope. Instructions for filling out the survey, including the matching elements, were reviewed for reading level using the Flesch–Kincaid formula and edited until they were under a fourth-grade reading level (Kincaid et al., 1975). Additional details about the survey administration process are available in the published study protocol (Agley et al., 2020).

**Selection of Self-Generated ID Code Elements**

This study used a seven-element matching procedure for fourth-grade students and an eight-element matching process for seventh- and 10th-grade students. As previously articulated, the matching elements were selected with the goal of maximizing the reality and perception of participant anonymity (e.g., not using elements derived
from parents’ names, respondents’ names, or birthdays). Unfortunately, up-to-date meta-analytic research had not identified SGICs that successfully met this criterion (Audette et al., 2020). Thus, researchers used procedures and principles from computer science and informatics literature on security access questions for forgotten passwords (e.g., Bonneau et al., 2015; Rabkin, 2008). These principles included applicability (likely to apply to all participants), unambiguity (only one correct answer), and memorability (salient to the population).

Certain elements from traditional SGICs met these standards and were not highly personalized, and so these were retained for this SGIC, but were unlikely to be sufficient for accurate matching (Audette et al., 2020). These were gender, race, ethnicity, and number of older siblings.

In a study of password challenge question creation, Just and Aspinall (2009) noted the efficacy of first-time events in developing security challenge questions, and their relative power in contrast to “favorite,” “best,” or other preference-based questions. Their study included consideration of a question about “first pet’s name,” which is unlikely to be known to the public or even to some friends and family, but it is likely to be remembered (Just & Aspinall, 2009). It is superior to a question about a “current” pet because the first pet is a fixed event (e.g., one might have had three pets as a child, but only one of them was first), and it is superior to a question about “favorite” pet because that might change over time. In the only other study of which we were aware that attempted a highly anonymous SGIC like ours (Wilson et al., 2010), the “favorite pet” element was problematic. We therefore selected an element for first pet’s name, and also included instructions to write the word “None” if there was no first pet in an attempt to disambiguate individuals who simply did not answer the question from those who affirmatively indicated they did not have a first pet.

Unfortunately, even setting aside potential concerns with memorability, other first-event type questions fail the applicability test for school-based data collection, such as “first job,” “make/model of first car,” or “where you met your spouse” (Schechter et al., 2009), simply because youth have had less time and opportunity to accumulate memorable first events. However, Schechter and colleagues also briefly shared other potential questions, such as “eye color.” We were conceptually concerned about that question because of the limited number of response options and high prevalence of a single eye color (brown), but identified that a similar functionality might be achieved by asking about backpack color, as most students carry backpacks, but there is some variability in coloration, and students likely can confirm their backpack color while answering the question.

Finally, some password retrieval mechanisms have attempted to use fixed numbers (e.g., “library card”) as a potential option for users, but these can be problematic because an individual may know the answer, or be able to retrieve it, but not at the time the question is being asked (Reeder & Schechter, 2011). However, many youths attending school have access to a locker that is secured using a combination of numbers. Those numbers have the unique characteristic of needing to be utilized every time a student wants to access their locker, theoretically improving the
likelihood of accurate recall (both memorability and unambiguity). Thus, we used the final two digits of the student’s locker combination as the final matching element. For this study, the following elements were used to generate an SGIC:

- Classroom number [preassigned by classroom, not entered by student]
- Gender [male or female]
- Race [White, African American or Black, Asian, Native American/Alaskan Native, Hawaiian/Pacific Islander, more than one race, other]
- “Are you Hispanic or Latino/a?” [yes or no]
- “How many older brothers and sisters do you have?” [0, 1, 2, 3, or more]
- “What color is your backpack? (if it’s several colors, what color is it mostly)” [black, red, green, blue, brown, pink, purple, a different color, I don’t have a backpack]
- “What was the name of your first pet? (if you have never had a pet before, write the word ‘None’ here)” [handwritten entry]
- “What are the last two digits in your school locker combination? (e.g., if your combination is 5-13-27, you would put 27) [two-digit entry from 0 to 9 each, I don’t have a lock].

**Computerized Matching Procedure (Step 1)**

A computer program was written in C# to perform the first step in the matching procedure. A numeric weight was assigned to each matching element based on its expected ability to discriminate between cases within a small cluster (classroom). In general, a match between elements with fewer response options (e.g., ethnicity, with two) is more likely than a match between elements with many response options (e.g., final two digits of school locker combination, with 100), and so weights were assigned in inverse relationship to the number of available response options. In absence of an established standard for this procedure, for this validation study, the following weights were used:

- 6: first pet’s name
- 3: backpack color, final two digits of school locker combination
- 2: gender, race, number of older siblings
- 1: ethnicity

An additional weight of 100 was assigned to classroom ID because all matches should logically occur within a classroom, and this weight prevented matches between classrooms. Using an oversized classroom weight—rather than only running the program within classrooms—allowed the program to identify potential matches with differing classroom IDs that may have resulted from miscoding or a smudged back of the response form. Each match had a summed score between 0 and 119 (seventh and 10th grades) or 0 and 116 (fourth grade) based on the elements that
matched within a computer-identified pair. In addition, single Levenshtein distance gaps (Black, 1999) in the pet name were allowed and assigned a penalty of 0.1, thereby allowing for minor errors in spelling or transcription (e.g., “Lilly” versus “Lily”) but scoring those cases differently in the output. Within-classroom matches had a minimum score of 100. The program operated as follows:

1. It calculated the score for each theoretically possible pair (e.g., for 1,571 pretests and 1,236 posttests, there were 1,941,756 possible pairs).
2. It sorted those pairs by score and iterated through them from the highest to lowest scoring, performing one of two actions:
   a. If neither the pretest nor posttest was in a previously accepted match, it accepted this pair as a match and saved it to a list.
   b. Otherwise, it rejected this pair as a match and discarded it.
3. This list of accepted matches was sorted by classroom and exported to a Microsoft Excel document for manual inspection and match data cleaning.

This process was similar to one based on within-code Levenshtein distances proposed by Schnell et al. (2010) and later used by Vacek et al. (2017), with the addition of weighting.

**Manual Matching Procedure (Step 2)**

The lead author conducted a line-by-line review of all proposed matches by examining the pretest data, posttest data, and computerized matching output simultaneously. Matches were reviewed iteratively by classroom. SGIC matching is improved when used in smaller clusters or subsamples (Faden et al., 2004), for example, by allowing manual review to apply logical rules beyond the script’s limitations. As an exemplar, in a classroom of 15 students, if there was only one pretest and one posttest that answered “zero” to older siblings, that answer’s uniqueness gave it greater power to determine a match than was assumed by the computer program, which only assigned it a weight of 2.

In addition, manual review allowed for decisions not programmed into the initial matching script (though future iterations of the program could feasibly incorporate these procedures), such as understanding that the following entries are likely the same for the first pet’s name: “Gary” and “my first was a fish named gary.”

In conducting this review, the lead author performed the following steps, starting with the highest score within a classroom:

1. Ensure absence of anomalies in classroom assignment that might affect computerized matching and correct the data set.
2. Review the matched pair with the highest score (or tied for the highest score) and either accept (confident or tentative) or reject the match.
3. Review the matched pair with the next-highest score and either accept (confident or tentative) or reject the match and proceed iteratively.

4. After reviewing all proposed matches, cross-validate all unmatched and tentatively matched cases to identify matches not proposed by the computer program.

The number of possible matches within a classroom was computed as the lower of $X$ and $Y$ below, because the upper limit of matches that can be identified is the lowest number of response forms at either response point:

- $X = (\text{No. of pretests distributed}) - (\text{No. of forms returned blank})$
- $Y = (\text{No. of posttests distributed}) - (\text{No. of forms returned blank})$

The match rate $[(\text{Matches})/(\text{Possible Matches})]$ was then computed overall and separately for each grade level. To determine whether any differences existed between grade levels, we used pairwise chi-square tests to compare proportions of successfully matched surveys by grade.

**Results**

**Overall Matching Results**

Data collection occurred from November 5, 2019, through March 11, 2020. The trial collected 1,571 response forms at pretest and 1,236 response forms at posttest. Importantly, the preponderance of this discrepancy ($n = 275$) resulted from the collision of the end of data collection with the initial orders to close schools due to the COVID-19 (coronavirus 2019) pandemic and the resulting inability to follow up with 11 classrooms at a single large school.

In total, the number of possible matches for the project was 1,184, which was the sum of possible matches for all classrooms (see the supplemental material available online for a Microsoft Excel file demonstrating computation of this value). A total of 931 matches were identified, yielding an overall match percentage of 78.63%. Of those matches, 856 were considered confident matches and 75 were considered likely but tentative matches during manual review. For the 65 classrooms for which matching was possible, matching percentages ranged from a low of 46.15% ($n = 1$) to a high of 100% ($n = 8$). The mean match percentage by classroom was 78.76% ($SD = 14.19\%$) and the median was 80.00%.

Table 1 illustrates the rejection ratios by computerized matching scores. After manual review, it turned out that all computed matches with a matching weight $> 113$ had been accepted as confident matches.

**Matching Results by Grade Level**

Seventh-grade students produced the highest accepted match rate (82.38%), which was significantly higher than both the fourth grade (71.03%; $\chi^2 = 9.3, p = .002$) and
Table 1. Crosswalk of Computerized and Manual Matching Outcomes.

| Computerized match score | No. of assignments<sup>a</sup> | % of Total pairs<sup>b</sup> | No. rejected | % Rejected at score | No. accepted as confident | No. accepted as tentative |
|--------------------------|-------------------------------|----------------------------|--------------|---------------------|--------------------------|--------------------------|
| 100–104.9                | 46                            | 4.5                        | 46           | 100.0               | 0                        | 0                        |
| 105–105.9                | 19                            | 1.9                        | 16           | 84.2                | 0                        | 3                        |
| 106–106.9                | 17                            | 1.7                        | 16           | 94.1                | 1                        | 0                        |
| 107–107.9                | 38                            | 3.7                        | 32           | 84.2                | 4                        | 2                        |
| 108–108.9                | 49                            | 4.8                        | 36           | 73.5                | 10                       | 3                        |
| 109–109.9                | 29                            | 2.8                        | 13           | 44.8                | 8                        | 8                        |
| 110–110.9                | 136                           | 13.3                       | 47           | 34.6                | 67                       | 22                       |
| 111–111.9                | 36                            | 3.5                        | 1            | 2.8                 | 24                       | 11                       |
| 112–112.9                | 40                            | 3.9                        | 1            | 2.5                 | 37                       | 2                        |
| 113–119                  | 614                           | 60.0                       | 0            | 0.0                 | 607                      | 7                        |
| NA<sup>c</sup>           | 195                           | —                          | 80           | —                   | 98                       | 17                       |

<sup>a</sup>The sum of number of assignments is 1,219, which is higher than the number of possible matches (1,184). This is because the computer assigned scores for 1,024 potential pairs. The remaining 195 database rows that were examined manually included some rows that were invalid and so were guaranteed to be rejected.

<sup>b</sup>Computed out of 1,024 computerized matches. These were primarily classrooms with issues with the scannable form itself (e.g., classroom ID 79 and 80 switched on posttest form codes generated on reverse of response sheet, but not on other documentation). All such cases were manually matched.
10th grade (76.52%, $\chi^2 = 5.5, p = .019$). The fourth-grade and 10th-grade match rates were not significantly different ($\chi^2 = 1.8, p = .180$).

### Percentages of Errors by Matching Element

We examined the individual matching elements within the accepted matches to determine the percentages of matches for which a given element was not the same. This process was like the one used by Audette et al. (2020). Importantly, these errors included both full element mismatches and instances where one or both elements in an accepted match were missing (except for locker number, as not all students have lockers in general, and fourth-grade students were not asked this question). These error rates, except for the rate for first pet’s name, are provided in Table 2.

Pet name could not easily be parsed into this data table. Normally, string-to-string comparison can be accomplished using matching procedures with error tolerance based on a Levenshtein distance, which counts the number of modifications (e.g., changing an “I” to an “E,” or deleting a letter) required to achieve a match between strings (Black, 1999). Our computerized script used a permissible Levenshtein distance of 1, which was likely too conservative. While many pet names matched exactly, spelling, transcription of handwriting, and extraneous detail (e.g., writing whether the pet was a dog or cat, including multiple pets) made string to string comparison by machine useful but not a comprehensive means of assessing this variable. For example, the certainty within a small cluster that “mordie cooper jasper” is the same as “Cooper, Mordie, and Jasper” is nearly 100%, but the Levenshtein distance between the strings is 16, which would be outside any parameter that would reasonably enable the program to match most names accurately.

However, it is important to note that the uniqueness of names made this matching element a valuable tool that was often individually sufficient to determine a match. Repetition of first pet names across all cases was fairly infrequent—the most common first pet name, by far, was “Max,” with only 19 instances across 1,571 pretests, or an average of one pet named “Max” per four pretest classrooms. As an exemplar of matching by pet name during manual review, consider Case 835, to which the

| Element                                                      | Error rate (%) |
|--------------------------------------------------------------|----------------|
| Gender                                                       | 3.7            |
| Ethnicity                                                    | 9.1            |
| Race                                                        | 9.5            |
| No. of siblings                                              | 14.6           |
| Backpack color                                               | 16.9           |
| Final locker combination number (excludes 103 fourth-grade cases) |                |
| [Exact Numeric Match] Only                                   | 59.9           |
| [Exact Numeric Match] or [Both Locker Numbers Missing = Match] | 31.2           |

Table 2. Error Rates for Matching Elements.
A computer program assigned a score of 105 because only gender, ethnicity, race, and classroom matched.

835: Gender(2), Ethnicity(1), Race(2), # Siblings(0), Backpack(0), Locker(0), First Pet’s Name(0; Lucky, she was a pitbull // Lucky)

In a classroom with relatively homogenous gender, race, and ethnicity, this match would not be feasible to make on the basis of those sociodemographic variables alone, and no other variables (No. of siblings, backpack, or locker) matched. However, in this small cluster of 17 cases with no other instances of name strings like “Lucky,” we were tentatively comfortable accepting this match as valid with the addition of the name.

**Discussion**

This was the first study, to our knowledge, to link adolescent participant data longitudinally using an SGIC completely devoid of personalized data elements extracted from names, birthdays, or addresses. A prior study attempted a similar procedure with returning U.S. veterans but encountered substantial issues with matching, likely due to element ambiguity (Wilson et al., 2010). Our study was also one of the first studies to use principles from password security questions to generate an SGIC for longitudinal data matching. A similar procedure was tested by Xu et al. (2020) and published while this study was ongoing, but their work was focused on matching optimization rather than perceptions of anonymity (e.g., they included the new variables as additional, not replacement, elements), and it was conducted in a laboratory setting among older adolescents.

We were able to match 78.63% of cases overall, with lower match rates for fourth and 10th grades, and a higher rate for seventh grade; this was consistent with our published expectation in the study protocol, which was 80% (Agley et al., 2020). We do not have any evidence indicating why there may be a discrepancy in matching success between the seventh and 10th grades, though fourth grade was expected to be lower because of the loss of a matching element (last number in the locker combination). As a proof of concept, these findings demonstrate that it is eminently possible to match data from adolescent participants with a high degree of anonymity and without using any personalized data elements, at least in the short term (2 weeks). It is, however, difficult to determine where this match rate falls within the spectrum of published SGIC match rates in terms of efficacy, as we outline subsequently.

**Comparison With Other Match Rates**

Audette et al. (2020) found that a substantive portion of the literature on SGIC matching cannot be mined effectively because insufficient details are available. That research team identified a wide range of “one-off” matching rates for different numbers of elements; the rates for seven and eight elements (the numbers used in this
study) in that review were 84.2% and 75.7%, respectively. It is difficult to ascertain what this means, though, as the numbers of studies used to calculate each mean varied (14 and 3, respectively) and the range of months for instances of follow-up ranged from 1 to 20. The highest match rates for SGICs tend to appear in (a) older studies, such as seminal work by Kearney et al. (1984), who reported a 92% matching success rate; (b) studies using numerous personal elements, such as Ripper et al. (2017), who reported a 95.7% two-off matching rate but used eight elements including birth month and year; and (c) the recent study by Xu et al. (2020), which suggests that it may be possible to achieve a very high success rate by combining traditional SGIC elements with newer elements, such as those used in this study. Each of these findings, though, is for SGICs that include personal elements.

Comparison between our match rate (78.63%) and any extant rate from the literature is substantively limited by the fact that ours was one of only two studies to exclude personal elements, the other being Wilson et al. (2010), which reported a 25% matching success rate at 3 months. Conceptually, there were also other factors that likely improved our rate, and some that likely suppressed our rate. As Audette et al. (2020) demonstrated, matching rates tend to decline over time, and so our relatively short time frame likely improved our rate by limiting data loss due to failed recall. At the same time, our computation of matching rate was conservative in how it handled missing data elements; we required all matching elements to be missing (e.g., a fully blank survey) to remove a survey from the denominator, meaning that mostly blank instruments that were “functionally” unmatchable were included in our computation and lowered our matching success rate. Many other studies with higher matching success rates have removed surveys with more than one missing element (e.g., Yurek et al., 2008); have been in contexts that allow more oversight of survey completion, such as a laboratory (e.g., Xu et al., 2020); or have been conducted using online data collection platforms that do not allow progression without complete data entry (e.g., Ripper et al., 2017).

We also speculate that our rate may have been suppressed because of how we treated instances of nonresponse. Studies using SGICs that sample at the level of individuals rather than clusters (i.e., classrooms) tend to compute matching rates after excluding those who do not respond (e.g., DiLorio et al., 2000), which may improve matching success rates relative to ours. We posit this theory because, with survey procedures that occur during the school day within classrooms, some individuals who do not want to respond might provide spurious matching elements rather than passing back a blank survey. For example, in our study, there were 16 reported locker combination numbers starting with the digit “6.” Improbably, 50% \((n = 8)\) of those numbers were “69,” only one of which was part of a matched case pair.

**Additional Factors Potentially Affecting Match Rates**

Notably, there were also eight classrooms within our study for which the matching success rate was 100%. On further examination, we observed that none of these
classrooms were fourth-grade classrooms. Furthermore, while the research team could not be involved with the survey procedures, one or more members often were present to conduct fidelity checks for the intervention itself, and so could observe survey administration from a distance. Six of the eight classrooms with perfect matching were at the same two schools, and we anecdotally observed that the instructors at these schools had excellent command of their classrooms relative to other participating schools. While not a definitive finding, this mirrors work just published by Palmer et al. (2020). Their team found that SGIC matching among college students was higher among students who also followed the protocols for other, non-SGIC components of the study. Thus, we speculate that strict adherence to the printed study protocols, and teacher involvement at all levels of survey administration, may have contributed to higher SGIC matching rates for classrooms in our study.

**SGIC Elements**

Among matched cases, our study found that, in general, more discriminating data elements had lower success rates (higher error rates). For example, gender was the same for 96.3% of matched cases (3.7% error), but backpack color, with nine possible responses, was only the same for 83.1% of matched cases (16.9% error). Two of the elements with intermediate numbers of possible responses (number of older siblings and race) had intermediate error rates, and also had relevant comparator data from prior research by Audette et al. (2020), who found a mean error rate of 16.4% for number of older siblings (ours was 14.6%), and a mean error rate of 13.5% for race (ours was 9.5%). Logically, the tradeoff for the increased number of possible responses for an element was increased confidence that a pretest and posttest were completed by the same person, given a match for that element. The clearest example of this phenomenon was first pet’s name, for which we did not directly compute an error rate, but which, to some degree, performed a superordinate matching function at the manual level, given the functionally infinite number of possible responses.

The last two digits of the locker combination were situated as an intermediary point between first pet’s name (with infinite possible responses) and the remaining elements, which had comparatively fewer response options. Matching locker numbers, with 100 possible responses, added meaningfully to confidence in a match. However, fewer students than expected had lockers or remembered their locker combinations accurately. Based on this study, we tentatively suggest that the last two digits of the locker combination may not be an efficient SGIC element. It was the same in relatively few (40.1%) matched cases, increasing to 68.8% when missing data at both points in time was considered a matched element (which approximated not having a locker, though not with 100% certainty). It is unclear what other elements might effectively replace it within school-based studies while still maintaining the desired reality and appearance of anonymity. One potential option might be asking about the first sport team the student joined, or the first sport played. Though it would not be a universally applicable element, 71.8% of youth aged 6 through 12
years played at least one day of a sport during that year, with reasonable variability in the specific sport played (The Aspen Institute, 2019). This SGIC element would have fewer possible responses than the last two digits of the locker combination, but more than the other elements we used (except first pet’s name), and would likely be applicable to more students, including those in elementary school.

Expanding on the “first instance” theme, there is a potential to incorporate multiple string-level variables when translating this mechanism to adults. Many of the password recall questions that were not salient for the school-based population, or that would likely have excluded large numbers of students, might be effective for adults—such as city where you had your first kiss. Importantly, though, memorability must still be taken into account (Rabkin, 2008), as some password recall questions that incorporate “first” or “unique” events or information that may not have high levels of memorability for the average person, such as last name of your third-grade teacher. Ambiguity must also be minimized in constructing such questions (Schechter et al., 2009)—such as by not simply asking where did you have your first kiss, which could produce a variety of correct locations. Use of multiple string-level variables such as these would not have been practical prior to the current computing era, but, as noted in the next subsection, SGICs no longer need to be limited to single or double characters per element.

**Matching Procedures and Approaches to Error Tolerance**

This study used a different approach to matching and error tolerance than most prior research, though there has not been a universal or consistent approach to error tolerance in SGIC matching. SGICs typically have taken the form of a compiled string such that each element is linked; for example, a code with middle initial, birth month, and gender might appear as D08M. For reasons such as user error and duplication (e.g., short SGIC strings in a large sample may legitimately be the same for multiple individuals), researchers have used techniques such as “one-off” matching, where all exact matches are processed, and then all matches that are the same except a single element are processed (e.g., Yurek et al., 2008). Later studies have applied a Levenshtein distance to the SGIC itself (e.g., Vacek et al., 2017), especially with longer SGICs, such that all exact matches are processed, then all one-off matches, then two-off matches, continuing up to a researcher-selected tolerance; for example, Vacek and colleagues provided matching rates iteratively up to four-off matches. Xu et al. (2020) suggest, and we agree, that there is no longer a computational need to compile SGIC elements into a string, and doing so can result in imprecision, especially in cases where an element is being truncated, as in “first two letters of your first pet’s name.” Collecting full rather than truncated and compiled SGIC data allows for certain elements like “first pet’s name” to introduce a high level of variability due to the number of possible entries (e.g., all possible names), but also, on matching (especially within a small cluster), to present a high degree of confidence in the match.
We opted to collect full SGIC elements, but unlike prior work, the primary difference between our computerized procedure and other recent approaches was inclusion of a weight for each element. The conceptual issue with procedures like “one-off” matching is that they assume that each element lends equal weight to the confidence that two SGICs match, which is not the case. As we describe, we assigned greater weights to elements with higher possible variability, producing matched scores between 100 and 119 (recall that classroom number had a matching weight of 100, since students could not match outside of their own classroom). Different from other processes, we also did not specify an a priori cutoff score where something was considered a match, instead opting to manually review each item for logical consistency. The manual review was completed iteratively by classroom and was subjective in the sense that it involved a researcher applying logical judgment to reject or accept cases. However, the weighted computerized matching procedure greatly constrained this review by providing suggested matches for cases, along with the weighted estimate. Thus, the primary decisions made during manual review related to matching first pet’s name strings that had the same meaning but used different strings (e.g., the prior example with a fish named Gary), or identifying errors with the forms themselves (e.g., a match score of exactly 19 indicating that a classroom number was likely incorrect, subject to confirmation). Interestingly, when cases were recompiled, there was a very clear delineation of scores, where no cases were rejected at a score of 113 or higher, two were rejected at scores between 111 and 112.9, and the rejection percentage jumped substantially (from 2.8% to 34.6%) when dropping from the [111–111.9] bracket to the [110–110.9] bracket. No cases below a score of 105 were identified as matching.

Limitations

This study had several limitations. First, data were collected only from students in fourth, seventh, and 10th grades, so the ability to generalize beyond these grade levels is limited, though inclusion of one grade from each level, elementary through high school, allows some inference of generalizability. Second, manual inspection and matching verification, while important, is a subjective process, even when following preexisting logical rules. To ensure full transparency, our study appendices include both the raw data for matching and documentation of exactly which cases were matched manually, and which cases were considered tentative matches (whether via computerized or manual match). In addition, several classrooms were matched entirely manually, but the mean rate of matching success at the classroom level was within 0.13% of the rate at the individual level, so we infer that the impact of this procedure was limited. Third, because the data specifically reference elements that pertain to adolescents in school (e.g., last digits of locker number, backpack color), these SGIC elements should not be used for individuals not enrolled in school. In addition, those same elements are likely static within, but not between, academic years. However, principles such as the matching approach (weighted computerized
matching and then manual inspection) and the use of a first pet’s name likely are transferrable to other settings, as described above. Fourth, we cannot know the exact causes of failed matches, but any discrepancy in actual school attendance, within a classroom, on the days of the pretest and posttest would affect the match rate denominator (e.g., knowing that 12 students took the pretest and posttest in a classroom doesn’t guarantee that they are the same 12 students). For longer term use of this SGIC method, movement between classrooms and schools may also occur at certain points, such as between breaks. Fifth, the fact that teachers administered the survey rather than researchers may also have affected matching rates, and potentially data quality, as we cannot know the extent to which teachers adhered to the instructions or were able to manage their classrooms. Finally, this study did not collect data directly assessing participants’ perception of study anonymity, so we cannot definitively assert that participants were more comfortable with these elements than with those extracted from personalized information.

Conclusion

In response to a research need, our team designed and tested a matching procedure for school-based adolescent longitudinal data collection that did not use any personalized data elements. In reviewing matching data from a large cluster, randomized trial, we identified that our approach was feasible and produced acceptable match rates in the short term. While our observed matching rate likely could have been improved by the addition of such elements, such as birth month, which had only a 1.5% mean error rate in Audette et al.’s (2020) meta-analysis, each additional personalized element may increase the risk of participant or observer concern with the study (Gassman et al., 2019). School and community support for research, even when it is benign and/or potentially beneficial with few possible detrimental outcomes, may rest not only on actual adherence to ethical conduct of research but also on affirmatively creating an appearance that this is so (where it may not be obvious; Gassman et al., 2016; Gassman et al., 2019). Additional research to identify other usable but nonpersonal longitudinal matching elements would be valuable and could potentially extend to other venues and populations such as military personnel, where confidentiality is key but prior attempts to link data have struggled (Wilson et al., 2010).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The ACT OUT! study, from which data were used, was funded by the Lilly Endowment, grant number 2019 0543.
Supplemental Material

Supplemental material for this article is available online.

Data Availability

Data used to accomplish matching are available in full on Figshare. A separate document summarizing matching ratios for this study is available as an appendix.

References

Agley, J., Jayawardene, W., Jun, M., Agley, D. L., Gassman, R., Sussman, S., Xiao, Y., & Dickinson, S. L. (2020). Effects of the ACT OUT! Social Issue Theater program on social-emotional competence and bullying in youth and adolescents: Protocol for a cluster randomized controlled trial. JMIR Research Protocols, 9(4), e17900. https://doi.org/10.2196/17900

Anderman, C., Cheadle, A., Curry, S., Diehr, P., Schultz, L., & Wagner, E. (1995). Selection bias related to parental consent in school-based survey research. Evaluation Review, 19(6), 663-674. https://doi.org/10.1177/0193841X9501900604

The Aspen Institute (2019). State of play: Trends and developments in youth sports. https://assets.aspeninstitute.org/content/uploads/2019/10/2019_SOP_National_Final.pdf

Audette, L. M., Hammond, M. S., & Rochester, N. K. (2020). Methodological issues with coding participants in anonymous psychological longitudinal studies. Educational and Psychological Measurement, 80(1), 163-185. https://doi.org/10.1177/0013164419843576

Black, P. E. (1999). Levenshtein distance. https://xlinux.nist.gov/dads/HTML/Levenshtein.html

Bonneau, J., Bursztein, E., Caron, I., Jackson, R., & Williamson, M. (2015, May). Secrets, lies, and account recovery: Lessons from the use of personal knowledge questions at Google [Paper presentation]. 24th International Conference on World Wide Web, Florence, Italy. https://doi.org/10.1145/2736277.2741691

Caruana, E. J., Roman, M., Hernández-Sánchez, J., & Solli, P. (2015). Longitudinal studies. Journal of Thoracic Disease, 7(11), E537-E540.

Coffelt, T. A. (2017). Confidentiality and anonymity of participants. In M. Allen (Ed.), The SAGE encyclopedia of communication research methods (pp. 227-230). Sage.

Colias, M. (2014, November 13). After complaints, school district postpones behavioral surveys. Rapid City Journal. https://rapidcityjournal.com/news/local/after-complaints-school-district-postpones-behavioral-surveys/article_415b37e1-9b69-5894-8d97-e41e71044478.html

Dent, C. W., Galaif, J., Sussman, S., Stacy, A., Burton, D., & Flay, B. R. (1993). Demographic, psychosocial and behavioral differences in samples of actively and passively consented adults. Addictive Behavior, 18(1), 51-56. https://doi.org/10.1016/0306-4603(93)90008-W

Dilorio, C., Soet, J. E., Van Marter, D., Woodring, T. M., & Dudley, W. N. (2000). An evaluation of a self-generated identification code. Research in Nursing & Health, 23(2), 167-174. https://doi.org/10.1002/(SICI)1098-240X(200004)23:2<167::AID-NUR9>3.0.CO;2-K
Durant, L. E., Carey, M. P., & Schroder, K. E. E. (2002). Effects of anonymity, gender, and erotophilia on the quality of data obtained from self-reports of socially sensitive behaviors. *Journal of Behavioral Medicine, 25*, 439-467. https://doi.org/10.1023/A:1020419023766

Faden, V. B., Day, N. L., Windle, M., Windle, R., Grube, J. W., Molina, B. S., Pelham, W. E. Jr., Gnagy, E. M., Wilson, T. K., Jackson, K. M., & Sher, K. J. (2004). Collecting longitudinal data through childhood, adolescence, and young adulthood: Methodological challenges. *Alcoholism, Clinical and Experimental Research, 28*(2), 330-340. https://doi.org/10.1097/01.ALC.0000113411.33088.FE

Fear, N. T., Seddon, R., Jones, N., Greenberg, N., & Wessely, S. (2012). Does anonymity increase the reporting of mental health symptoms? *BMC Public Health, 12*, Article 797. https://doi.org/10.1186/1471-2458-12-797

Fradette, K., Keselman, H. J., Lix, L., Algina, J., & Wilcox, R. R. (2003). Conventional and robust paired and independent-samples $t$ tests: Type I error and power rates. *Journal of Modern Applied Statistical Methods, 2*(2), 481-496. https://doi.org/10.22237/jmasm/1067646120

Gassman, R. A., Agley, J., Fly, A. D., Beckmeyer, J. J., He, K., Sayegh, M. A., & Tidd, D. (2016). Using biological samples for youth ATOD survey validation. *Addiction Research & Theory, 24*(3), 177–185. https://doi.org/10.3109/16066359.2015.1094061.

Gassman, R. A., Dutta, T., Agley, J., Jayawardene, W., & Jun, M. (2019). Social media outrage in response to a school-based substance use survey: Qualitative analysis. *Journal of Medical Internet Research, 21*(9), e15298. https://doi.org/10.2196/15298

Helms, A. D. (2020, February 22). CMS yanks controversial survey questions about student sexual orientation, gender ID. *WFAE 90.7*. https://www.wfae.org/post/cms-yanks-controversial-survey-questions-about-student-sexual-orientation-gender-id#stream/0

Hetzel, L., Boivin, J.-M., Di Patrizio, P., & Chau, K. (2019). Assessment of participation biases for a confidential non-anonymous adolescent study: A population-based study. *Psychiatry Research, 271*, 30-36. https://doi.org/10.1016/j.psychres.2018.12.154

Hoeprner, J. E. (2017). “You need to shut up”: Research silencing and what it reveals about academic freedom [Unpublished doctoral dissertation]. Australian National University.

Johnson, T. P. (2014). Sources of error in substance use prevalence surveys. *International Scholarly Research Notices, 2014*, Article 923290. https://doi.org/10.1155/2014/923290

Just, M., & Aspinall, D. (2009, July). Personal choice and challenge questions: A security and usability assessment [Paper presentation]. 5th Symposium on Usable Privacy and Security, Mountain View, CA, United States. https://doi.org/10.1145/1572532.1572543

Kearney, K. A., Hopkins, R. H., Mauss, A. L., & Weisheit, R. A. (1984). Self-generated anonymous collection of longitudinal questionnaire data. *Public Opinion Quarterly, 48*(1B), 370-378. https://doi.org/10.1093/poq/48.1B.370

Kelley, D. (2015, September 6). Controversial Healthy Kids survey back in Colorado schools this fall. *The Gazette*. https://gazette.com/business/controversial-healthy-kids-survey-back-in-colorado-schools-this-fall/article_f6d932bb-763f-5e61-a4d9-8600c296d917.html

Kemper, J. (2008). The chilling effect: How to researchers react to controversy? *PLOS Medicine, 6*(1), e1000014. https://doi.org/10.1371/journal.pmed.1000014

Kincad, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel* (Report No. DP0008946). University of Central Florida. https://doi.org/10.21236/ADA006655
Malvin, J. H., & Moskowitz, J. M. (1983). Anonymous versus identifiable self-reports of adolescent drug attitudes, intentions, and use. *Public Opinion Quarterly, 47*(4), 557-566. https://doi.org/10.1086/268812

NBCWashington. (2014, November 8). Parents discuss controversial school sex survey. https://www.nbcwashington.com/news/local/parents-discuss-controversial-school-sex-survey_washington-dc/1975972/

O’Malley, P. M., Johnston, L. D., Bachman, J. G., & Schulenberg, J. (2000). A comparison of confidential versus anonymous survey procedures: Effects on reporting of drug use and related attitudes and beliefs in a national study of students. *Journal of Drug Issues, 30*(1), 35-54. https://doi.org/10.1177/00220426000300103

Palmer, J. E., Winter, S. C., & McMahon, S. (2020). Matching anonymous participants in longitudinal research on sensitive topics: Challenges and recommendations. *Evaluation and Program Planning, 80*, Article 101794. https://doi.org/10.1016/j.evalprogplan.2020.101794

Rabkin, A. (2008, July). Personal knowledge questions for fallback authentication: Security questions in the era of Facebook [Paper presentation]. 4th Symposium on Usable Privacy and Security, Pittsburgh, Pennsylvania, United States. https://doi.org/10.1145/1408664.1408667

Reeder, R., & Schechter, S. (2011). When the password doesn’t work: Secondary authentication for websites. *IEEE Security & Privacy, 9*(2), 43-49. https://doi.org/10.1109/MSP.2011.1

Ripper, L., Ciaravino, S., Jones, K., Jaime, M. C. D., & Miller, E. (2017). Use of a respondent-generated personal code for matching anonymous adolescent surveys in longitudinal studies. *Journal of Adolescent Health, 60*(6), 751-753. https://doi.org/10.1016/j.jadohealth.2017.01.003

Schechter, S., Brush, A. J. B., & Egelman, S. (2009, May). It’s no secret: Measuring the security and reliability of authentication via “secret” questions [Paper presentation]. IEEE 30th Annual Symposium on Security and Privacy, Berkeley, CA, United States. https://doi.org/10.1109/SP.2009.11

Schnell, R., Bachteler, T., & Reiher, J. (2010). Improving the use of self-generated identification codes. *Evaluation Review, 34*(5), 391-418. https://doi.org/10.1177/0193841X10387576

Seifert, T. A., Pascarella, E. T., Erkel, S. I., & Goodman, K. M. (2010). The importance of longitudinal pretest-posttest designs in estimating college impact. *New Directions for Institutional Research, 2010*(S2), 5-16. https://doi.org/10.1002/ir.368

Shaw, T., Cross, D., Thomas, L. T., & Zubrick, S. R. (2014). Bias in student survey findings from active parental consent procedures. *British Educational Research Journal, 41*(2), 229-243. https://doi.org/10.1002/berj.3137

Spence, S., White, M., Adamson, A. J., & Matthews, J. N. (2015). Does the use of passive or active consent affect consent or completion rates, or dietary data quality? Repeat cross-sectional survey among school children aged 11-12 years. *BMJ Open, 5*(1), e006457. https://doi.org/10.1136/bmjopen-2014-006457

Vacek, J., Vonkova, H., & Gabrhelik, R. (2017). A successful strategy for linking anonymous data from students’ and parents’ questionnaires using self-generated identification codes. *Prevention Science, 18*(4), 450-458. https://doi.org/10.1007/s11121-017-0772-6

White, V. M., Hill, D. J., & Effendi, Y. (2004). How does active parental consent influence the findings of drug-use surveys in schools? *Evaluation Review, 28*(3), 246-260. https://doi.org/10.1177/0193841X03259549
Wilson, A. L. G., Hoge, C. W., McGurk, D., Thomas, J. L., Clark, J. C., & Castro, C. A. (2010). Application of a new method for linking anonymous survey data in a population of soldiers returning from Iraq. *Annals of Epidemiology*, 20(12), 931-938. https://doi.org/10.1016/j.annepidem.2010.08.008

Xu, S., Chan, A., Lorber, M. F., & Chase, J. P. (2020). Using security questions to link participants in longitudinal data collection. *Prevention Science*, 21(2), 194-202. https://doi.org/10.1007/s11121-019-01080-8

Yurek, L. A., Vasey, J., & Havens, D. S. (2008). The use of self-generated identification codes in longitudinal research. *Evaluation Review*, 32(5), 435-452. https://doi.org/10.1177/0193841X08316676