Design and Application of Self-Generated Identification Codes (SGICs) for Matching Longitudinal Data

J. Direnga
Research Assistant, Engineering Education Research Group
Hamburg University of Technology
Hamburg, Germany
E-Mail: julie.direnga@tuhh.de

D. Timmermann[1]
Research Assistant, Engineering Education Research Group
Hamburg University of Technology
Hamburg, Germany
E-Mail: dion.timmermann@tuhh.de

J. Lund
Student
Hamburg University of Technology
Hamburg, Germany
E-Mail: dion.timmermann@tuhh.de

C. Kautz
Professor for Engineering Education, Engineering Education Research Group
Hamburg University of Technology
Hamburg, Germany
E-Mail: kautz@tuhh.de

Conference Key Areas: Engineering Education Research
Keywords: subject-generated coding, matching algorithm

INTRODUCTION

Educational sciences investigate learning of individuals and groups and the effectiveness of courses, where learning is a process that can be interpreted as the change in scores on appropriate test instruments [1, 2]. In cases where the baseline is unknown, this requires at least two measurements and individuals must often be tracked as part of longitudinal studies [3–6]. One way to accomplish this tracking is the usage of official

[1] Corresponding Author
D. Timmermann
dion.timmermann@tuhh.de

identification codes, e. g. matriculation numbers. However, in cases where protection of personally identifiable information is of interest, official identification codes cannot be used. In these cases, so-called self-generated identification codes (SGICs), also referred to as “subject-generated coding” [7], are an alternative [8–11]. Judged by the number of publications found on SGICs, they seem to be commonly used in the medical and social sciences, while they are less frequent in the educational sciences.

With the use of SGICs, individuals can easily and consistently compose their identification code based on a set of coding questions provided to them. For successful data linkage, these coding questions must reliably result in the same code string for an individual over the course of the investigation (accuracy), and must furthermore result in different codes for different individuals (identifying power). Additionally, they should protect the individual’s anonymity by not allowing anyone to identify the individual based on the code itself. To accomplish these goals, the coding questions used for an SGIC should follow certain criteria which will be discussed below.

The Engineering Education Research Group at Hamburg University of Technology has collected over 5000 pre- and about 3500 post-tests in different university courses over the past 12 years. Until 2015, we used matriculation numbers to match the tests. As matriculation numbers are official and unique identification numbers, one could expect high matching rates. However, we were only able to match 72 % of the post-tests. There are three aspects that might explain this low rate: (1) there can always be errors in writing or reading the matriculation number, (2) students are sometimes *deliberatly* providing missing or wrong information, and (3) there is no guarantee that every student on the post-test has also taken part in the pre-test and vice versa. Using SGICs, we do not expect to reduce the effect of aspects (1) and (3), but rather the effect of aspect (2), i. e. we could get more students to trust us with their data. Any matching rate equal or better than 72 % is therefore acceptable.

Since switching to SGICs, we have collected over 1000 pre- and 700 post-tests. In this paper, we will describe the criteria used for the selection of coding questions and report our experience with these criteria. Additionally, we will describe the matching algorithm used by our group. The intent of this paper is to present our results and experiences in order to help other researchers in the community to decide whether SGICs might be useful for them.

1 SELF-GENERATED IDENTIFICATION CODES (SGICS)

SGICs are composed of several items. Each item is an answer to a coding question. In related publications, items are also referred to as variables [7], elements [9, 10] or components [12]. They are pieces of information drawn from attributes inherent to the individual, such as day or place of birth, mother’s first name, etc. The individual can thus reliably compose his or her individual code by answering the coding questions anywhere and at any time. At the same time, he or she cannot easily be identified by others through the code itself [8, 9, 12]. Investigating the perceived respondent burden and anonymity through using the SGICs, Damrosch found that “subjects saw the code as easy to generate ([mean] M = 4.96 for the 0 to 5 scale) and difficult to ‘break’ (M= 4.85); subjects also were satisfied with the protection of their anonymity (M= 4.89)” [12]. Carifio and Biron report statements from high school students, which allow for a similar conclusion [13]. One disadvantage of SGICs, however, is that the codes are not necessarily unique, i. e. there may exist two or more individuals in the population under investigation with identical codes. The following section will discuss criteria that should be considered when choosing SGIC items.

1.1 Criteria for SGIC items

The publications on SGICs referenced in this work all name some criteria that the items must fulfill. However, there seems to be no exhaustive list that the literature agrees upon. The following will elaborate on the criteria we used for selecting our set of items. We have divided our list into two sections: criteria that SGIC items must fulfill and criteria that SGIC items should fulfill. For each criterion we will give one example where it is not fulfilled.

Each SGIC item must...

- ... apply to every person. (e. g. not "name of first pet", as not everyone has had a pet)
- ... be well known by the individuals [10, 12, 14] but not by the researcher [12]. (e. g. not "blood group")
- ... not change over time [7, 9, 12, 14]. (e. g. not "number of siblings", as the parents may have a another child in the future)

SGIC items should...

- ... be uniquely identifiable [14]. (e. g. not "hair color" as this can already become quite difficult to say reliably with natural hair colors. People who have dyed their hair might also be confused which color to pick.)
- ... be an unobservable attribute of the person [12]. (e. g. not "sex", especially in small groups)
- ... be something people are willing to say. (e. g. not "PIN code of mobile phone", as a person might not want to disclose this information)
- ... have a high variation [7, 9] among the group so it is helpful in discriminating. (e. g. not "current academic affiliation", when the individuals are currently all enrolled in the same university)
- ... be simple to state/understand (e. g. not "fifth letter in your mother's maiden name or last letter, if name consists of less than five letters"). This is also important to avoid sampling bias [8, 9], e. g. not only the participants with high cognitive ability should be able to answer consistently.

Apart from the criteria mentioned above, there are two important statistical measures when selecting items for SGICs. These are their accuracy, i. e. the probability that the same person will answer the coding question identically at each code generation event, and their identifying power, which is related to the probability that two different individuals have the same value for an item [10].

1.2 Length of the code

Finding the optimal number of items thus results in a trade-off between the accuracy versus the identifying power where the size of the population is another important parameter because it influences the identifying power. Having too few items, decreases the identifying power, whereas having too many items (1) decreases the accuracy of the entire code and (2) increases the respondent burden. Damrosch found that the respondents' acceptance of their eight-item code was very good with respect to respondent burden [12] Based on these results, we choose to use eight items or less for our code. As seven-item codes were also often used in literature with populations of similar sizes [7, 9, 10] and we found seven items that sufficiently comply with the criteria mentioned above, we decided on the code(s) described in the following section.

1.3 SGIC items used by our group

Figures 1, 2, and 3 show the three versions of SGICs that were used by our group. We decided to arrange the items such that letters and numbers alternate for better readability. The first five-item SGIC, shown in Figure 1, was used in our first trial run. As we knew beforehand that the cohort would only consist of about 100 students, we anticipated that five items would provide sufficient identification power. The seven-item code shown in Figure 2 was introduced for a larger cohort of about 300 students. The same code was used in a different format for an even larger cohort of about 600 students on a bubble sheet shown in Figure 3. The advantages and disadvantages of the handwritten and the bubble-sheet formats will be discussed in Section 1.4.

The SGIC items that we chose largely fulfill the criteria mentioned above. Nevertheless, some items might be problematic for some individuals. Kearney et al. report "relatively high error rates in number of older siblings and father's initial (perhaps due to changing family composition)" [9] and suggest not to use these items. The error rates reported by Schnell et al. confirm these results only concerning the number of older siblings, but not concerning the father's first name [10]. We believe that these error rates strongly depend on the population and that we can safely use these items in our context, especially in combination with a matching algorithm that tolerates off-1 matches (see Section 2).

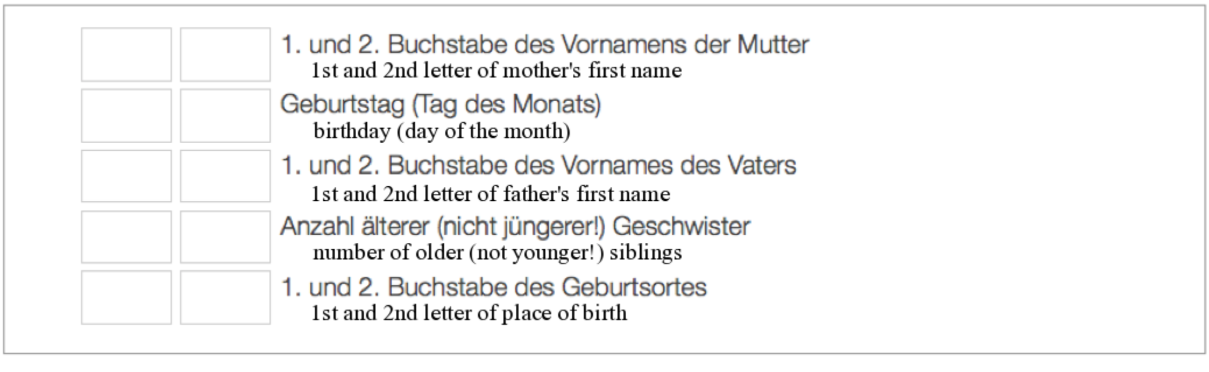
The number-of-older-siblings item does not have a high variation by design. Few people in our investigated population have more than three siblings and the necessity of asking only for older siblings reduces this number even further. We could gain more variation by splitting the siblings up into brothers and sisters. The resolution of the information of "2 older siblings" can then be tripled by saying "1 older brother and 1 older sister", "2 older brothers and 0 older sisters" or "0 older brothers and 2 older sisters". [2]

1.4 Code forms

Figures 2 and 3 show the same code in different formats, a handwritten and a machine-readable bubble-sheet version. Both formats have advantages and disadvantages which will be discussed below. Generally, we tried to reduce the respondent burden by reducing the amount of text for coding questions and instructions or examples to a minimum.

Bubble-sheets have several advantages over handwritten codes. Damrosch for example used a form with pre-printed response options "[t]o avoid any difficulty with illegible

[2] We are aware that, due to e. g. transgender issues, the value of the item might change over time for some individuals, however, we chose to ignore this problem because we do not expect significant case numbers.



□ □ 1. und 2. Buchstabe des Vornamens der Mutter
1st and 2nd letter of mother's first name

□ □ Geburtstag (Tag des Monats)
birthday (day of the month)

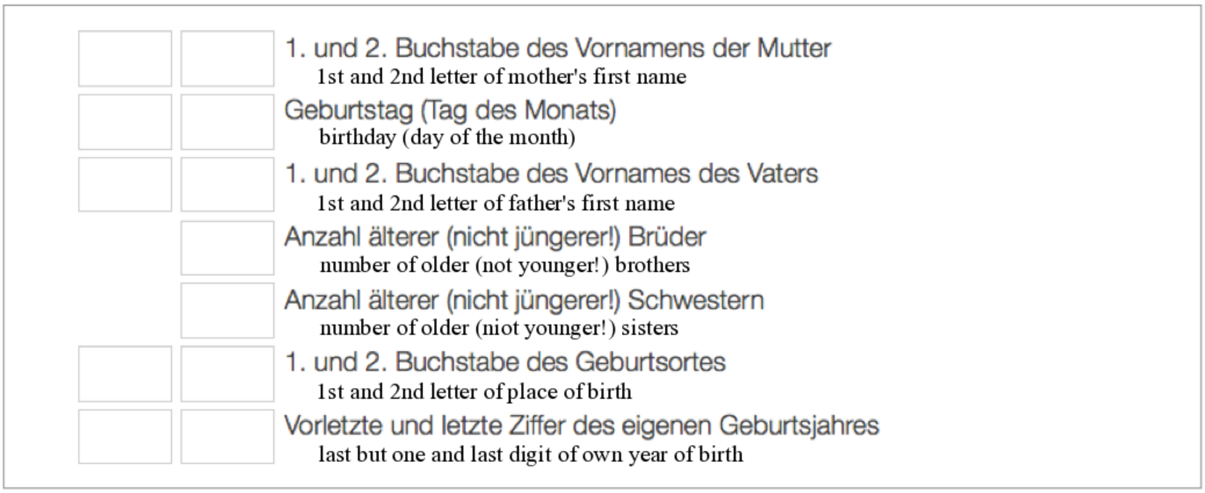
□ □ 1. und 2. Buchstabe des Vornames des Vaters
1st and 2nd letter of father's first name

□ □ Anzahl älterer (nicht jüngerer!) Geschwister
number of older (not younger!) siblings

□ □ 1. und 2. Buchstabe des Geburtsortes
1st and 2nd letter of place of birth

Fig. 1. Five-item (preliminary) handwritten version of the SGIC. The english translation of the coding questions was added for this publication.

handwriting” [12]. Furthermore, the bubble-sheets provide a limited possible answering option, and thereby can help to clarify the coding question. For example, an attempt was made to clarify the question “own birthday” by adding the phrase “(day of the month)”. In the handwritten version, some students wrote what seems to be the first two letters of their birth month, not the day of the month. In the bubble-sheet version, this error is less likely to occur, as the answering options 1 to 31 indicate that the answer should be a day. In the next revision of the bubble-sheet, we will most likely remove the verbal clarification “(day of the month)”, as we have reason to suspect that the term “month”



□ □ 1. und 2. Buchstabe des Vornamens der Mutter
1st and 2nd letter of mother's first name

□ □ Geburtstag (Tag des Monats)
birthday (day of the month)

□ □ 1. und 2. Buchstabe des Vornames des Vaters
1st and 2nd letter of father's first name

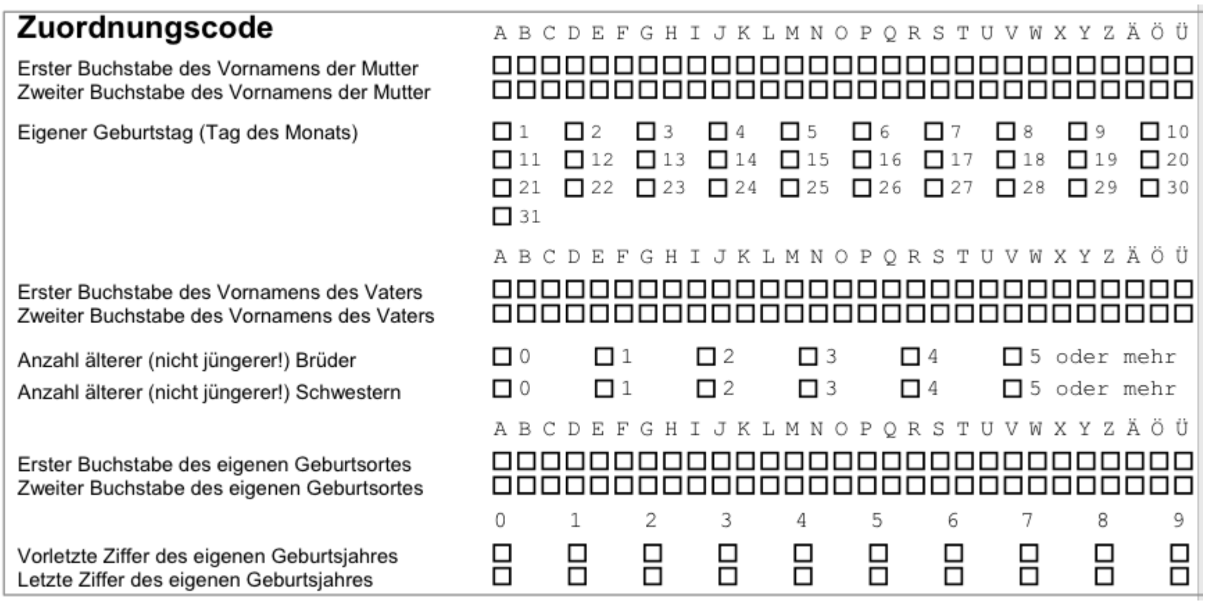
□ Anzahl älterer (nicht jüngerer!) Brüder
number of older (not younger!) brothers

□ Anzahl älterer (nicht jüngerer!) Schwestern
number of older (niot younger!) sisters

□ □ 1. und 2. Buchstabe des Geburtsortes
1st and 2nd letter of place of birth

□ □ Vorletzte und letzte Ziffer des eigenen Geburtsjahres
last but one and last digit of own year of birth

Fig. 2. Seven-item (final) handwritten version of the SGIC. The english translation of the coding questions was added for this publication.



Zuordnungscode

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Ä Ö Ü

Erster Buchstabe des Vornamens der Mutter □□□□□□□□□□□□□□□□□□□□□□□□□□□□□
Zweiter Buchstabe des Vornamens der Mutter □□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Eigener Geburtstag (Tag des Monats)
□ 1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □ 8 □ 9 □ 10
□ 11 □ 12 □ 13 □ 14 □ 15 □ 16 □ 17 □ 18 □ 19 □ 20
□ 21 □ 22 □ 23 □ 24 □ 25 □ 26 □ 27 □ 28 □ 29 □ 30
□ 31

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Ä Ö Ü

Erster Buchstabe des Vornamens des Vaters □□□□□□□□□□□□□□□□□□□□□□□□□□□□□
Zweiter Buchstabe des Vornamens des Vaters □□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Anzahl älterer (nicht jüngerer!) Brüder □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 oder mehr
Anzahl älterer (nicht jüngerer!) Schwestern □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 oder mehr

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Ä Ö Ü

Erster Buchstabe des eigenen Geburtsortes □□□□□□□□□□□□□□□□□□□□□□□□□□□□□
Zweiter Buchstabe des eigenen Geburtsortes □□□□□□□□□□□□□□□□□□□□□□□□□□□□□

0 1 2 3 4 5 6 7 8 9

Vorletzte Ziffer des eigenen Geburtsjahres □ □ □ □ □ □ □ □ □ □
Letzte Ziffer des eigenen Geburtsjahres □ □ □ □ □ □ □ □ □ □

Fig. 3. Seven-item bubble-sheet version of the SGIC. The coding questions translate as follows: “first letter of mother’s first name”, “second letter of mother’s first name”, “own birthday (day of the month)”, “first letter of father’s first name”, “second letter of father’s first name”, “number of older (not younger!) brothers” [options limited to 5 or more], “number of older (not younger!) sisters” [options limited to 5 or more], “first letter of own place of birth”, “second letter of own place of birth”, “last but one digit of own year of birth”, “last digit of own year of birth”.

could actually trigger the error mentioned above. Finally, an obvious advantage is that the bubble-sheet can more easily be evaluated for large sample sizes.

A clear disadvantage is that bubble-sheets only allow certain answers. When asking for the first letters of names, for example, all possible characters have to be provided for the participants to select, which might make the code difficult to fill in for students whose parents' names contain special characters.

On the other hand, the handwritten version is more intuitive to fill in. Instead of having to find the correct bubble to mark, one simply writes down the code. We saw several cases where two bubbles in the same row were marked and the next row was blank, making it impossible to tell in which order the characters should be in the code string. All the bubble-sheet options also require more space, this might make it difficult to fit the code form and the test questions on the same page.

2 MATCHING SGICs

Once a good set of items has been chosen and the codes have been collected by appropriate code forms, the data stemming from the same individuals must be linked. As reported in the literature, matching rates can significantly increase by also accepting matches that differ in one or more items (e. g. [7, 9], otherwise "losses up to 50 % of the cases are not unusual" [10]. At the same time, allowing for errors increases the risk of incorrectly linking data. The maximum number of varying items discussed in the examined literature is two (e. g. [11]), although most come to the conclusion, that applying an "off-1"-technique yields the best results with respect to avoiding false-positive matches and at the same time finding more true-positive pairs (e. g. [7–9]). Kearney et al. report the improvement of their matching rates by accepting off-1 matches. They achieved "92 % linkage of cases over a one-month interval and 78 % over a one-year interval" [9] as opposed to 46 % exact matches for long-term, and 67 % for short-term. The reported rate of incorrect (false-positive) matches is less than 2 %.

Schnell et al. [10] propose a more sophisticated matching algorithm with a greater error tolerance. They make use of the Levenshtein distance, which also allows for errors in the order of the items. In case of our bubble-sheet code form, this could be quite useful (see Section 1.4). As it is otherwise quite unlikely that students switch the order of items by mistake, and this matching technique requires a greater number of items and might therefore result in increased "respondent burden" and "higher nonresponse rates" [10], we decided to use a simpler algorithm, described below.

2.1 Algorithm used for matching

We implemented a very simple matching algorithm which is based on the concept of distances. For each post-test code string it calculates the distance to all pretest code strings, where the distance between two strings is the number of SGIC items that are different (not the number of characters!). Missing items are always counted as different, even if two missing items are compared. These values are written to a table with one row per post-test and one column per pretest. The algorithm searches for entries in the table that are smaller than all other values in both their respective column and row (see Table 1). These are possible pairs of SGICs because they have the smallest distance to each other among all other possible combinations. If there is a predefined error tolerance limit, i. e. matches may differ in at most n items, the minimum distance must be smaller or equal to n. Non-unique pre- or post-tests will never be matched, as identical tests will always have the same distance to every other test and thus can never have the smallest distance to any test.

Table 1. Example for pre- and post tests and the distances of their respective code strings. The highlighted code strings are identical and have a distance of 0.

Pre / Post	AN02KL11LU94	SA12UL00RE96	BI18JA00LA88	BI22ST00LE96
BI18JA00LA99	7	5	1	4
AN03KL11LU94	1	7	7	7
BI18JA00LA88	7	6	0	5

Table 1 shows a minimal example: looking from row 1, the SGIC in this row would match to the SGIC in column 3. However, when looking from column 3, the SGIC in column 3 would have the minimum distance to the SGIC in row 3 and not the one in row 1. Therefore, the SGICs in row 3 and column 3 match, as long as there is no other minimum in row 3, which is the case here. Once a match has been identified, the pair is removed from the table. Due to the sequential nature of the algorithm, some matches are only found after another iteration over all yet unmatched codes.

2.2 Error tolerance level

The error tolerance level corresponds to the maximum distance tolerated in the matching process. Exact matches will be referred to as off-0 matches, if one item may be different and the match shall still be accepted, it is called an off-1 match etc.

The chosen tolerance level is a trade-off between receiving more false-positive or false-negative results and it should depend on the specific setting. False-positive results (i. e. codes from different individuals have been matched by the algorithm) can have a negative effect on the validity of the experiment. When the data is analysed for individuals and small sample sizes, false-positives need to be avoided by all means, whereas few cases are acceptable when aggregated data is analysed and the sample sizes are large. False-negatives are less severe. As long as there is no matching bias, they only reduce the effective sample size.

2.3 Accuracy of the algorithm

Some related studies have a “reference” [10] or “gold standard” [7] independent from the SGIC, from which the true matches can be determined. We do not have this possibility in our case. Instead, as suggested by Kearney et al. [9], we used comparison of handwriting to observe the frequency of false-positives and false-negatives generated by our matching algorithm. According to Schnell et al. [10], comparison of handwriting was often used as an indicator in other studies and thus seems to be an accepted method. Of the 314 hand-written post-tests in our database, 220 could be matched by the algorithm to the pre-tests. Except for one case where the analysis of handwriting was inconclusive, we can clearly state that none of these matches are false-positive, i. e. pairs of tests that erroneously were considered a match by the algorithm. The number of false negatives was only 17, i. e. only 17 post-tests could be matched using the handwriting analysis, but did not fulfill the criteria of the algorithm. This is only 8 % of all hand-written post-tests. As the SGICs collected on the bubble-sheets are matched by the same algorithm, we consider this to be the upper limit of false-negatives.

With the new algorithm and this particular SGIC schema, we were able to match 76 % of the 715 post-tests collected in 2015 and 2016 including off-1 matches. The matching interval, i. e. the time between pre- and posttest was about 3 months in both cases.

3 DISCUSSION

The matching rate of 76 % including off-1 matches is similar to results found in previous studies (e. g. [7,9,10]). Other studies even report matching rates of about 90 % (e. g. [13]). In any case, comparing our matching rates to the matching rates of other studies is not very useful, as the conditions are often different. Instead, we limit our discussion to comparing our achieved matching rates of matriculation numbers on the one hand and SGICs on the other hand.

Even though matriculation numbers are unique, using them as matching criterion did not yield more matches in total compared to the SGICs. On the contrary, with SGICs, we have achieved a slightly higher matching rate (76 %) than with official matriculation numbers (72 %). Our initial goal was to achieve at least a comparable matching rate.

As discussed in Section 1.4, the coding question ”own birthday” is not simple enough to understand in the handwritten version of the code form. As it is a very good item with regard to all other criteria on the list, we decided to keep it in the schema. However, there is room for improvement.

4 SUMMARY

In this article, we presented a list of criteria for high quality SGIC-items and discussed the advantages and disadvantages of our chosen items for the purpose of pre-/post-testing and possible application in long-term study designs. We compared a handwritten and a bubble-sheet version of code forms and presented the algorithm used for matching. With accepting off-1 matches, we could improve our matching rate to 76 % compared to 72 % matching via official matriculation numbers. A handwriting analysis indicates no false-positive and only 8 % false-negative matches.

Based on our results, we conclude that the advantages of using SGICs can outweigh the disadvantages, providing that the items and the matching algorithm are chosen carefully. We thus encourage other researchers to consider using SGICs for their data linkage.

REFERENCES

[1] Hake, R. R. (1998) Interactive - engagement versus traditional methods: A six - thousand - student survey of mechanics test data for introductory physics courses. *Am. J. Phys.*, **66**, 64.

[2] Hestenes, D. and Wells, M. (1992) A Mechanics Baseline Test. *The Physics Teacher*, **30**, 159–166.

[3] Dellwo, D. R. (2010) Course assessment using multi-stage pre/post testing and the components of normalized change. *Journal of the Scholarship of Teaching & Learning*, **10**.

[4] Pollock, S. J. (2009) Longitudinal study of student conceptual understanding in electricity and magnetism. *Physical Review Special Topics-Physics Education Research*, **5**, 020110.

[5] Pawl, A., Barrantes, A., Pritchard, D. E., and Mitchell, R. (2012) What do seniors remember from freshman physics? *Physical Review Special Topics - Physics Education Research*, **8**, 020118.

[6] Bond, L. (2005), Carnegie Perspectives: Who has the lowest prices? *retrieved from* `http://archive.carnegiefoundation.org/perspectives/who-has-lowest-prices`, accessed on: 2016-07-28.

[7] McGloin, J., Holcomb, S., and Main, D. S. (1996) Matching anonymous pre-posttests using subject-generated information. *Evaluation review*, **20**, 724–736.

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

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

[10] Schnell, R., Bachteler, T., and Reiher, J. (2010) Improving the Use of Self-Generated Identification Codes. *Evaluation Review*, **34**, 391–418.

[11] DiIorio, C., Soet, J. E., Van Marter, D., Woodring, T. M., and Dudley, W. N. (2000) An evaluation of a self-generated identification code. *Research in nursing & health*, **23**, 167–174.

[12] Damrosch, S. P. (1986) Ensuring anonymity by use of subject-generated identification codes. *Research in nursing & health*, **9**, 61–63.

[13] Carifio, J. and Biron, R. (1978) Collecting Sensitive Data Anonymously: The CDRGP Technique. *Journal of Alcohol and Drug Education*, **23**, 47–66.

[14] Hogben, L., Johnstone, M. M., and Cross, K. W. (1948) Identification of medical documents. *British medical journal*, **1**, 632.