CENTRAL EUROPEAN REVIEW
OF ECONOMICS AND MANAGEMENT

ISSN 2543-9472; eISSN 2544-0365

Vol. 9, No.4, December 2025, 49-68

www.cerem-review.eu
www.ojs.wsb.wroclaw.pl

A Privacy-Preserving Method for Longitudinal Participant Linkage in Web Surveys

Rafał PALAK

Wrocław University of Science and Technology, Poland

ORCID: **https://orcid.org/0000-0002-4632-7709**
E-mail: **rafal.palak@pwr.edu.pl**
Received: 14.12.2025, Revised: 24.122025, Accepted: 28.12.2025
doi: http:// 10.29015/cerem.1043

Aim: To enable longitudinal linkage in online panel surveys without collecting direct identifiers and while aligning with modern data-protection requirements

Design / Research methods: The article proposes a client-side protocol where participants create a reproducible secret from a self-chosen pseudonym and an ordered image sequence. The browser normalizes and cryptographically hashes these inputs to derive a short alphanumeric core code, adds a modulus-97 checksum for strict local validation, and the backend stores only a salted hash scoped to a specific study (form-family) context.

Conclusions / findings: This paper introduces a client-side protocol for generating anonymous yet linkable participant identifiers in web-based surveys by deriving a reproducible code from a user-chosen pseudonym and image sequence entirely in the browser, and by storing only a form-family–salted hash on the server for longitudinal linkage within a study. The design incorporates a checksum for strict client-side validation and is intended to reduce spurious identifiers caused by typographical errors; empirical validation of matching performance, usability, and security properties is left for future work.

Originality / value of the article: The work refines SGIC-style respondent-generated linkage by combining graphical secrets with browser-based cryptographic processing, checksum-based client-side validation, and form-family salting-yielding a concrete, implementable algorithm that improves privacy-respecting longitudinal linkage.

Keywords: longitudinal survey methodology, anonymous respondent linkage, self-generated identification codes (SGIC), data privacy in empirical research
JEL: C81, C83.

Correspondence address: Rafał Palak Faculty of Information and Communication Technology, Wrocław University of Science and Technology, Poland. E-mail: rafal.palak@pwr.edu.pl

1. Introduction

Online surveys and web-based questionnaires have established themselves as central instruments in the social sciences, public health, and human-computer interaction research (Platje et al. 2025). Many of these studies are longitudinal by design, seeking to correlate measurements from the same individual across different time points or multiple instruments. At the same time, ethical and legal requirements are increasingly restricting the use of direct identifiers, such as names, email addresses, or institutional accounts, creating a tension between the need for stable linkage and the obligation to preserve anonymity.

A common compromise is the self-generated identification code (SGIC) (Yurek et al. 2008). In a typical SGIC scheme, participants construct a personal code based on answers to a set of stable questions (e.g., initials, birth dates, or family names) and re-enter this code in subsequent waves. While this approach avoids the creation of formal user accounts and the storage of explicit contact details, decades of empirical use have exposed structural limitations. Matching rates often fall substantially below 100%, and unmatched cases are rarely random; individuals who fail to reproduce their codes frequently differ systematically from those who succeed, introducing bias due to attrition. Furthermore, many standard SGIC "recipes" rely on quasi-identifiers that are increasingly scrutinized under modern privacy frameworks, such as the GDPR, due to their potential for re-identification.

Concurrent with these methodological challenges, the technical capabilities of web browsers have evolved. Modern browsers can now perform robust client-side processing, including the generation of random keys and the computation of cryptographic hashes. This capability offers an opportunity to shift identification logic from the server to the user's local environment. However, existing survey tools rarely fully exploit this potential. Yet most survey tools and methodological work on anonymous longitudinal linkage still assume textual SGICs or server-side pseudonymization starting from known identities, leaving a gap between human-friendly procedures that respondents can reliably repeat and technically robust, unlinkable representations suitable for privacy-aware data management.

This article proposes a novel client-side protocol for generating anonymous yet linkable participant identifiers. The method combines a user-chosen pseudonym with a user-selected sequence of images, fusing them into a single, memorable "secret." From this secret, a short alphanumeric core code is deterministically derived in the browser. A critical feature of this design is the inclusion of a cryptographic checksum, which enables strict client-side validation. This mechanism significantly reduces the probability that typographical errors or partial recall will generate spurious new identifiers, which is a common failure mode in traditional SGICs.

The identifier ultimately transmitted to and stored by the survey platform is not the core code itself, but a salted cryptographic hash incorporating a project-specific "family salt." This ensures that responses can be reliably linked within a specific study while making it computationally infeasible to link identities across independent projects or to reconstruct the underlying secrets. The primary contribution of this paper is the formulation of a concrete, reproducible algorithm for this process. We specify the exact sequence of normalization, encoding, hashing, and checksum computation, moving beyond generic calls for "better anonymity" to offer an implementable standard. A secondary contribution is the theoretical analysis of the construction's properties, specifically how the hybrid approach (pseudonym + visual secret) addresses the trade-off between memorability and entropy in anonymous linkage. The family-level salt is intended to restrict cross-project linkability and limit re-identification risks. The emphasis remains on the algorithm's structure, information flow, and privacy characteristics rather than on empirical evaluation.

The remainder of this paper is organized as follows. The next section surveys the existing literature on anonymous longitudinal linkage. This section is followed by a detailed introduction to the proposed client-side algorithm. The subsequent section discusses the proposed approach. Finally, the last section concludes the paper and outlines directions for future work.

2. Literature review

Research on anonymous linkage in longitudinal surveys spans three distinct traditions: self-generated identification codes (SGICs), broader discussions of coding participants in anonymous studies, and more recent work on cryptographically inspired, client-side identifier generation. In this section, we review each of these strands and identify the gap that motivates the proposed algorithm.

2.1. Self-generated identification codes in longitudinal research

Self-generated identification codes (SGICs), also referred to as subject-generated identification codes, were introduced as a pragmatic solution for linking repeated measurements without requiring the collection of overt identifiers such as names or addresses. In the seminal study by Kearney et al., participants in school-based substance use surveys constructed codes from personal information (for example, initials and elements of dates), which were then used to match questionnaires across waves (Kearney et al. 1984). Kearney et al. reported relatively high matching rates-around 92% over a one-month interval and approximately 78% over a one-year interval-when using a combination of exact and relaxed matching rules. These early results established SGICs as a viable method for anonymous longitudinal linkage, particularly in adolescent health research. Subsequent studies examined SGICs in more detail and in different contexts. Grube et al. evaluated seven-element SGICs in a panel study of adolescent substance use and found that exact matching succeeded for roughly 71% of cases over a one-month period, with improved rates when near-matches (differing by one element) were accepted. DiIorio et al. conducted an evaluation of SGIC performance in a multi-wave study and documented that matching success declines with time between waves, while errors in specific code components (such as names of parents or the order of siblings) are particularly frequent. Yurek et al. conducted empirical evaluations of SGICs in a large, multi-wave longitudinal study of registered nurses, focusing on the reliability of anonymous linkage over 6-, 12-, and 18-month intervals. Their findings show that match rates ranged from roughly 50% to 67%, with most mismatches resulting from respondent errors in specific SGIC elements, particularly those referencing family members or other information with low personal salience (Yurek et al. 2008). The study demonstrates that even small inaccuracies in participant-generated code components can substantially reduce matchability and that unmatched cases may contribute to sample attrition, potentially leading to bias. However, the authors' propensity score analysis suggests that such bias was limited in their data. Importantly, the authors note that SGIC performance was markedly lower in heterogeneous adult organizational settings than in prior school-based studies, and they call for refined and more user-friendly approaches to anonymous linkage to improve match reliability in such contexts.

Prior work has repeatedly shown that SGIC-based linkage often suffers from substantial proportions of unmatched cases as high as 50%, motivating research into more robust and user-centered identification schemes (DiIorio et al. 2000). Direnga et al. provides a design-oriented overview of SGIC construction, specifying desirable properties of code elements (stability over time, memorability, low sensitivity, and sufficient variability) and illustrating how concrete SGIC recipes meet or fail to meet these criteria.

Beyond technical matching performance, several authors emphasize that unmatched cases are not random. Research by Grube et al. shows systematic differences between matched and unmatched respondents in behaviors and sociodemographic characteristics, indicating selection bias (Grube et al. 1989). DiIorio et al. similarly notes that individuals whose codes cannot be matched over time may differ in important ways from those with consistent codes. More recent methodological discussions of participant coding underline that SGIC-based linkage can introduce complex patterns of attrition and misclassification, particularly over longer follow-up intervals (Audette et al. 2020).

Taken together, this body of work demonstrates that SGICs are an established technique with clear strengths (no need for accounts and no explicit storage of names); however, they also exhibit structural weaknesses. Matching rates rarely reach 100%, decline with increasing time between waves, and unmatched cases often exhibit systematic differences compared to matched participants. Moreover, many SGIC recipes rely on quasi-identifiers, such as birth dates and parental names, which raise contemporary privacy concerns in light of data linkage capabilities and legal frameworks, including the GDPR (DiIorio et al. 2000).

2.2. Recent developments and critiques of SGIC approaches

Recent contributions revisit SGICs in light of contemporary data-protection standards and participants' perceived anonymity. Calatrava et al. (2022) argue that many traditional SGIC recipes rely on personalized elements—such as letters from parents' names or components of birth dates—that can render respondents identifiable under modern legal frameworks like the GDPR. In response, they propose a "fully private" SGIC built from non-personal, stable childhood preferences, aiming to reconcile strict anonymity requirements with sufficient longitudinal matching reliability. Brändle and Plaschke provides a contemporary assessment of the accuracy of self-generated identification codes (SGICs) using large-scale administrative and school-based assessment data. Their analysis quantifies the frequency of SGICs requiring error-tolerant matching and identifies individual and contextual factors associated with non-exact linkages (Brändle 2024). The findings reaffirm that SGICs remain a practical tool for anonymous data linkage in educational research; however, successful implementation depends on careful attention to subgroups and school environments, where SGICs are more prone to inaccuracies.

Methodological reviews of coding strategies for anonymous longitudinal research, such as the analysis by Audette et al. (2020), position self-generated identification codes (SGICs) alongside alternative approaches, including de-identified data, preexisting unique identifiers, and electronic anonymizing systems. Their review highlights the trade-offs that each method presents in terms of anonymity, participant trust, feasibility, and matching accuracy. SGICs are described as relatively straightforward for researchers to implement and as offering a high degree of perceived anonymity, but they rely on respondents' consistent recall and reporting of the information used to construct the code. As a result, some data loss is expected due to inaccuracies such as omitted or inconsistently reported code components.

More applied studies illustrate these challenges in concrete longitudinal settings. For example, feasibility studies linking anonymous data from children and adolescents in prevention research report moderate matching rates using SGIC-like identifiers and highlight difficulties related to cultural heterogeneity, family structure, and variable interpretations of code questions (Vacek et al. 2024). These findings suggest that, despite decades of use, SGIC designs still struggle to balance memorability, stability, and privacy across diverse populations.

2.3. Anonymous IDs, client-side generation, and short codes

In parallel to SGIC-focused work, a distinct line of research has emerged around anonymous participant identifiers generated algorithmically, often without requiring respondents to remember complex codes. Sandnes introduces CANDIDATE, a tool that generates short, anonymous participant IDs for multi-session studies. The tool operates locally in the browser, creating simple alphanumeric IDs with a low collision rate and strong anonymity properties. Simulations demonstrate that ID spaces of moderate size can support typical study sample sizes with negligible collision risk, provided that ID generation follows a suitably randomized scheme (Sandnes 2021a). Similar ideas are explored in HIDE, which proposes short IDs for robust and anonymous linking of users across multiple sessions in small HCI experiments, again emphasizing anonymity and low implementation overhead (Sandnes 2021b).

These tools demonstrate the feasibility of utilizing modern browser capabilities to generate participant-linking IDs without requiring server-side knowledge of secrets and without requiring respondents to construct textual SGICs. However, they are typically designed for relatively small studies or HCI experiments and do not explicitly address scenarios in which respondents must be able to reconstruct their identifiers across devices or waves themselves. Instead, IDs are often generated and managed by the researcher or the tool, and participants are expected to retain or re-use system-generated codes.

More broadly, the literature on privacy-aware record integration and pseudonymization has proposed a variety of techniques for linking records based on encrypted or hashed identifiers, sometimes involving trusted third parties or secure multi-party computation (Kuzu et al. 2013). While these approaches are powerful in institutional contexts (for example, linking administrative registers or clinical databases), they typically assume that a stable, trusted identifier exists from the outset (such as a patient number) and that organizations control the underlying infrastructure. They are less directly applicable to anonymous web surveys where respondents participate from personal devices and are not willing to expose any stable personal identifier.

2.4. Graphical secrets and memorability

A further relevant strand of work comes from research on graphical passwords and image-based authentication. Studies in this field repeatedly show that users tend to remember images and spatial patterns better than arbitrary alphanumeric strings, and that image-based secrets can be made both memorable and reasonably secure when designed carefully (Wiedenbeck et al. 2005). Although this literature focuses primarily on authentication in security-sensitive systems rather than research participation, its core insight–that visual choices can serve as user-generated, memorable secrets–suggests that image-based components could be integrated into SGIC-like schemes for survey linkage.

The research conducted by (Krótkiewicz et al. 2018; Wojtkiewicz et al. 2024) and his team highlights the importance of user-centered design in systems that involve ongoing user interaction, particularly concerning cognitive load modeling and interface efficiency. Their findings demonstrate that the reliability of measurements hinges on both the algorithms used and how tasks are presented. This conclusion supports the incorporation of image-based components and simplified input workflows in the current proposal, aiming to minimize user-induced linkage errors in online panel surveys.

Some methodological discussions of SGIC design already hint at the importance of cognitive load and memorability, arguing that code elements should be easy to recall and stable over time (DiIorio et al. 2000; Direnga et al. 2016). However, to date, there appears to be little explicit integration between graphical secret design and anonymous survey linkage: existing SGIC recipes remain almost entirely textual, and graphical methods are rarely discussed in the survey methodology literature.

2.5. Summary and motivation

The literature thus reveals three key observations. First, SGICs are a well-established and widely used method for anonymous longitudinal linkage, but they suffer from incomplete matching and potential selection bias. They often rely on quasi-identifiers that raise modern privacy concerns (DiIorio et al. 2000; Grube et al. 1989; Kearney et al. 1984). Second, recent algorithmic tools, such as CANDIDATE and HIDE, demonstrate that anonymous, short identifiers can be generated locally with strong guarantees of anonymity. However, these tools typically do not exploit user-memorable secrets, nor do they fully address respondent-side reconstruction of identifiers over time (Sandnes 2021a; Sandnes 2021b). Third, the potential of graphical secrets to improve memorability has been recognized in the security literature but has not yet been systematically applied to anonymous survey linkage.

This combination of findings suggests a methodological opportunity: to design an algorithm that preserves the core intuition of SGICs (respondent-generated, reproducible identifiers) while replacing quasi-identifying textual elements with a structured combination of pseudonyms and graphical choices, and embedding these in a modern client-side cryptographic construction. The proposed approach aims to occupy this space precisely, complementing existing SGIC designs and anonymous ID tools by specifying a concrete, browser-side algorithm for anonymous yet linkable identifiers in online survey research.

3. Proposed method

The proposed method defines a client-side algorithm for generating anonymous yet linkable participant identifiers in web-based surveys. Its central idea is to replace traditional textual self-generated identification codes with a structured, cryptographically processed secret composed of a user-chosen pseudonym and a user-selected sequence of images. All operations involving human-readable secrets are carried out entirely within the participant's browser. The survey backend receives and stores only a salted hash derived from this secret, which serves as the participant identifier for longitudinal linkage within a defined family of forms.

3.1. Participant' secret and normalization

Each participant provides two elements, namely a pseudonym P entered as free text and an ordered sequence of images $\boldsymbol{I} = (i_1, i_2, \ldots, i_k)$, $k \in \mathbb{N}$ selected from a fixed catalog in the survey interface (each i_j is an integer in a bounded range (for example, $0 \leq i_j \leq 63$). The first step of the construction is to map $(P, \boldsymbol{I})$ into a canonical intermediate representation. The pseudonym is processed by a normalization function $norm(\cdot)$ that trims leading and trailing whitespace, converts all characters to lowercase, replaces locale-specific letters with their basic Latin counterparts, and removes any remaining characters that are not letters or digits. This yields a normalized pseudonym

$$P^* = norm(P),$$

which ensures that trivial variations in spelling, case, or diacritics do not produce different identifiers.

3.2. Deriving the core code

The image sequence is encoded as a decimal string that uniquely represents both its length and its elements. A simple encoding function $enc(\cdot)$ can be defined as follows: the length k is written as a two-digit decimal number with leading zeros if necessary, and each image identifier i_j is also written as a two-digit decimal number. These substrings are concatenated to form a single string

$$S = enc(\boldsymbol{I})$$

For example, if $\boldsymbol{I} = (5, 27, 9, 42)$ and $k = 4$, then S would be the string "0405270942". This encoding is purely internal to the algorithm and is not exposed to the participant.

From P^* and S, the browser derives a short alphanumeric core code CORE that serves as an intermediate, human-readable representation of the participant's secret. A cryptographic hash function H (such as SHA–256) is applied separately to the normalized pseudonym and to the encoded image sequence, producing two hash values

$$h_P = H(P^*), h_I = H(S),$$

which are treated as hexadecimal strings. These two strings are concatenated with a delimiter (for example, the character "|") to form a combined string

$$C = h_P \| " - " \| h_I.$$

A second application of the hash function yields

$$h_C = H(C).$$

The value h_C is then mapped to a fixed-length alphanumeric code. To this end, the first m bytes of h_C (for example, $m = 5$, corresponding to 10 hexadecimal characters) are interpreted as a non-negative integer $n = bytesm(h_C)$. This integer is encoded in base 36 using digits and uppercase letters by a function $base36(\cdot)$, giving

CORE = base36(n),

which is padded with leading zeros or truncated as necessary to obtain a fixed length L (for example $L = 8$). The mapping

$$(P, \boldsymbol{I}) \to P^* \to S \to h_P\,, h_I \to h_C \to CORE$$

is deterministic for a given pseudonym and image sequence; however, assuming the cryptographic properties of *H*, it is computationally infeasible to invert and recover P or $\boldsymbol{I}$ from $CORE$.

3.3. Deriving the core code

To support the detection of typographical errors when participants enter their code manually, a checksum is computed over $CORE$ using a $modulus - 97$ construction similar to that employed in IBAN validation. Each character of $CORE$ is mapped to a numeric value in $\{0, \dots, 35\}$ by interpreting digits $0 - 9$ as $0 - 9$ and letters A–Z as 10–35. These values are concatenated as decimal strings to form a large integer *N* represented as a string. Two zeros are appended to this string, corresponding to multiplication by 100, and the remainder

$$r = N\,00\,mod\,97$$

is computed using an iterative modulus algorithm that operates on the decimal string to avoid overflow. The checksum is defined as

$$c = 98 - r,$$

with c set to 0 if $c = 98$, and it is then written as a two-digit decimal string in the range $"00" - "97"$. The human-readable version of the participant code that may be displayed to the respondent is

$$FULL = CORE"\text{-}"CHECKSUM$$

where $CHECKSUM$ is the two-digit representation of c. All of these operations occur entirely within the browser.

The checksum is used purely for client-side validation. When a participant chooses to work with the textual form of their code, the survey interface can request the pseudonym, the image sequence, and the full code they believe to be correct. The browser recomputes $CORE$ from the current $(P, \boldsymbol{I})$ pair, recalculates the checksum, and verifies that both match the user-entered $CORE - CHECKSUM$. If either the core or the checksum is inconsistent, the code is rejected locally, and the participant is prompted to correct the input. In this way, malformed or inconsistent identifiers are never transmitted to the server, reducing the creation of spurious new identifiers due to typographical mistakes.

3.4. Final identifier and linkage

The identifier used for linkage in the survey backend is obtained by hashing $CORE$ together with a project-specific secret, $FORM\_FAMILY\_SALT$, chosen by the survey administrator. This salt is a randomly chosen value by the survey administrator for a given family of forms (for example, all waves of a single study) and is kept secret on the server side. It ensures that the same underlying secret (P, I) does not lead to identical identifiers across distinct projects that employ different salts. Given a core

$$PARTICIPANT\_ID = H(CORE \| " - " \| FORM\_FAMILY\_SALT).$$

The result is a hash value that can be represented as a hexadecimal string or in another canonical encoding, depending on backend requirements. Only $PARTICIPANT\_ID$ (and optionally the checksum as a non-sensitive auxiliary field) is sent to the server together with the survey responses and is stored in the database; neither the pseudonym, the image sequence, nor the core code is transmitted or persisted.

Within the overall survey workflow, the method is used whenever a participant first enrolls or returns for a subsequent wave. In a typical scenario, the survey interface displays a text field for the pseudonym and a panel of images from which the participant selects their personal sequence. Upon submission, the browser computes P^*, S, $CORE$, the checksum, and finally $PARTICIPANT\_ID$ using the secret salt associated with the form family. The participant may be shown $FULL\_CODE$ for their own records, but the server receives only $PARTICIPANT\_ID$ and the answers to the survey questions. When the same participant returns in a later wave of the same study, they enter the same pseudonym and select the same images, the browser repeats the computation, and the resulting $PARTICIPANT\_ID$ matches the value stored for their previous responses. Longitudinal linkage is thus achieved solely by equality of salted hashes, without the survey platform ever learning or storing the human-readable secrets that generated them.

3.5. Discussion

The proposed algorithm aims to strike a balance between traditional textual SGIC schemes and fully system-generated anonymous identifiers. By combining a user-chosen pseudonym with a user-selected sequence of images and processing this composite secret entirely on the client, the method preserves the central intuition of SGICs-that respondents themselves provide the material needed for longitudinal linkage–while replacing quasi-identifying code elements such as dates of birth or parental initials with a more abstract, cryptographically mediated construction. This section discusses the methodological and privacy-related implications of this design, focusing on expected effects on matching performance, error profiles, anonymity, and practical deployment in web-based surveys.

From the perspective of matching performance, the most important change introduced by the algorithm is the separation between the human-facing secret and the machine-facing identifier. In classical SGICs, respondents must remember and re-enter the same textual recipe (often combining several elements), and any inconsistency directly affects the matching key stored on the server. In the proposed scheme, the only values that reach the server are salted hashes of the derived core code; the heavy reliance on user memory and consistency is shifted to the client, where it can be supported by validation mechanisms that prevent formed identifiers from being recorded. The cryptographic checksum on the core code precisely addresses the problem that typographical errors and partial recall lead to spurious new identifiers. As long as the pseudonym and image sequence are reproduced correctly, the client can detect inconsistencies between the recomputed core and the user-entered code and can reject invalid codes before any data is submitted. Conceptually, this should reduce the fraction of unmatched cases arising purely from input errors. However, the method does not eliminate unmatched cases resulting from genuine changes in the participant's secret or complete forgetting.

The use of images as part of the secret is motivated by cognitive considerations. Image-based secrets capitalize on the tendency of people to remember visual patterns and stories more readily than arbitrary alphanumeric strings, suggesting that a small sequence of images coupled with a pseudonym may be easier to recall over time than a multi-element textual SGIC. At the same time, the image sequence is never transmitted or stored; it is only used locally to derive a hash that participates in the construction of the core code. This design aims to leverage the memorability of graphical secrets without disclosing additional quasi-identifying information or rich metadata to the backend. However, it also introduces a new kind of dependency: the participant must be able to recognize the same image catalog and reconstruct their sequence in subsequent waves. If the visual design of the image panel changes significantly, or if different devices confusingly render images, reconstruction may fail, resulting in unmatched cases. The method, therefore, implicitly assumes a stable and well-designed image set throughout the study's life.

Regarding privacy and anonymity, the algorithm is deliberately conservative. The server never receives the pseudonym P, the normalized pseudonym P^* , the image sequence $\boldsymbol{I}$, the encoded image string S, or the intermediate core code $CORE$. It only receives a salted hash of $CORE$, where the salt is specific to a form family. Under standard assumptions about the cryptographic hash function, this means that inverting the participant identifier to recover the underlying secret is computationally infeasible, and that the same participant generates different identifiers in different projects that use various salts. This mitigates the risk of unintended cross-project linkage and constrains the consequences of a data breach: an attacker obtaining the database of identifiers would not be able to view any human-readable components and would lack the necessary salts to apply dictionary attacks effectively. The method thus aligns with the concept of pseudonymization at the interface between client and server, while aiming to achieve a level of protection that approximates anonymity from the server's perspective.

The introduction of a form family salt also has methodological implications. Because $PARTICIPANT\_ID$ is derived from both $CORE$ and $FORM\_FAMILY\_SALT$, identifiers are inherently scoped to the family of forms for which the salt is defined. This scoping is desirable when researchers wish to avoid linking individuals across independent studies or institutional boundaries; however, it also means that deliberate cross-study linkage (for instance, combining two projects conducted by the same team) cannot be achieved without reusing or coordinating salts at the time of data collection. The algorithm, therefore, encodes a particular stance on linkage: it makes within-study linkage easy and automatic, while making cross-study linkage a deliberate design decision rather than an accidental side-effect of code reuse. Despite these advantages, several limitations and potential sources of bias remain. The method cannot prevent unmatched cases that arise when participants intentionally change their pseudonyms or image sequences, or when they forget their secrets entirely. In such situations, the algorithm correctly treats the responses as coming from a new anonymous individual. Whether this behavior is desirable depends on the analytic goals: for some studies, a strict requirement that only genuinely reproducible secrets lead to linkage may be appropriate; for others, it may be preferable to tolerate more aggressive matching rules at the risk of occasional false links. Furthermore, the cognitive burden of choosing and remembering an appropriate pseudonym and image sequence may itself be associated with participant characteristics such as age, education, or digital literacy. If certain groups are more likely to choose unstable secrets or to forget them, then even a technically robust algorithm may still induce differential matching rates and selection effects.

The introduction of a form family salt also has methodological implications. Because $PARTICIPANT\_ID$ is derived from both $CORE$ and $FORM\_FAMILY\_SALT$, identifiers are inherently scoped to the family of forms for which the salt is defined. This scoping is desirable when researchers wish to avoid linking individuals across independent studies or institutional boundaries; however, it also means that deliberate cross-study linkage (for instance, combining two projects conducted by the same team) cannot be achieved without reusing or coordinating salts at the time of data collection. The algorithm, therefore, encodes a particular stance on linkage: it makes within-study linkage easy and automatic, while making cross-study linkage a deliberate design decision rather than an accidental side-effect of code reuse. Despite these advantages, several limitations and potential sources of bias remain. The method cannot prevent unmatched cases that arise when participants intentionally change their pseudonyms or image sequences, or when they forget their secrets entirely. In such situations, the algorithm correctly treats the responses as coming from a new anonymous individual. Whether this behavior is desirable depends on the analytic goals: for some studies, a strict requirement that only genuinely reproducible secrets lead to linkage may be appropriate; for others, it may be preferable to tolerate more aggressive matching rules at the risk of occasional false links. Furthermore, the cognitive burden of choosing and remembering an appropriate pseudonym and image sequence may itself be associated with participant characteristics such as age, education, or digital literacy. If certain groups are more likely to choose unstable secrets or to forget them, then even a technically robust algorithm may still induce differential matching rates and selection effects.

Security considerations also warrant discussion. The scheme protects against server-side misuse and external attackers, but does not prevent impersonation by someone who learns another participant's secret. If a participant discloses both their pseudonym and their image sequence to a third party, the third party can, in principle, reproduce the same CORE and therefore the same $PARTICIPANT\_ID$ in future waves of the same study. This limitation is shared with traditional SGICs and passwordbased systems: any method based on a user-controlled secret is vulnerable to voluntary disclosure. The algorithm reduces the risk of *involuntary* leakage by never transmitting the secret itself, but it cannot control what participants choose to reveal. In practical survey contexts, the incentives for such impersonation are typically low, but the possibility should be acknowledged when assessing the overall threat model. From an implementation standpoint, the method is intentionally modest in its technical requirements. It assumes the availability of a standard cryptographic hash function in the browser, basic string and numeric operations, and the ability to store and use a single form family salt on the server side. This approach makes integration into existing survey platforms conceptually straightforward: the client-side logic can be packaged as a reusable script, and the backend can treat $PARTICIPANT\_ID$ as just another string field used for grouping responses. At the same time, this simplicity means that the method does not attempt to implement more advanced privacy techniques such as differential privacy, secure multi-party computation, or hardware-backed key storage. It is best understood as a structured refinement of SGIC-like linkage tailored to the affordances of current web technologies, rather than as a comprehensive privacy framework.

Taken together, these considerations suggest that the proposed algorithm addresses several key weaknesses of classical SGICs—particularly the direct exposure of quasi-identifiers and the fragile coupling between user input and stored identifiers—while introducing its own set of assumptions about user behavior, interface stability, and salt management. Its main strengths lie in the clear separation between human-facing secrets and machine-facing identifiers, the use of client-side validation to reduce error-induced unmatched cases, and the scoping of identifiers to specific study families. Its principal vulnerabilities relate to participant memory, voluntary disclosure, and the potential for differential matching performance across subgroups. These trade-offs are inherent to the problem of anonymous longitudinal linkage and frame the conditions under which the method can be considered appropriate for a given research design.

5. Conclusions and future work

This paper introduces a client-side method for generating anonymous yet linkable participant identifiers in web-based surveys. The approach replaces the traditional SGIC-based linkage with a structured secret composed of a user-chosen pseudonym and an image sequence, which is processed locally through normalization, encoding, and cryptographic hashing. By exposing only, a salted hash of the derived core code to the server, the method supports longitudinal linkage while preserving strong anonymity and preventing the backend from learning any human-readable identifiers. The use of a checksum further strengthens reliability by reducing the likelihood that typographical errors produce spurious identifiers.

The proposed scheme demonstrates how principles from SGIC design, graphical secrets, and browser-based cryptography can be combined into a practical mechanism that aligns with modern privacy expectations. While the method offers more apparent separation between human-facing and machine-facing identifiers, it retains certain limitations: matching performance still depends on participants' ability to reproduce their pseudonym and image sequence, and the design remains vulnerable to intentional disclosure of the secret. As such, the method serves as a conceptual refinement of, rather than a complete solution to, all challenges posed by participant-generated identifiers.

Future work should validate the approach empirically, particularly its impact on matching rates, error patterns, and subgroup differences in usability. Additional research is needed to formalize the security properties of the scheme, assess resistance to guessing and brute-force attacks, and evaluate how parameters such as salt management and core length affect robustness. Finally, practical integration into survey platforms and user-interface studies will be essential to determine how participants understand, remember, and reliably reproduce the required secret.

In summary, the method provides a clear and technically grounded direction for improving anonymous longitudinal linkage, illustrating how client-side cryptographic processing can enhance both privacy protection and methodological reliability in survey research.

Acknowledgments

The author thanks the Department of Applied Informatics at Wrocław University of Science and Technology for their support. I acknowledge the use of ChatGPT for assistance in reviewing my writing and supporting the translation from my first language.

References

Audette L.M., Hammond M.S., Rochester N.K. (2020), Methodological issues with coding participants in anonymous psychological longitudinal studies, “Educational and Psychological Measurement”, vol. 80 no. 1, pp. 163–185.

Brändle T., Pläschke A. (2024), Beyond matching rates: examining the accuracy of self-generated ID codes, https://osf.io/preprints/osf/k98j6_v1 [21.12.2025].

Calatrava M., de Irala J., Osorio A., Benítez E., Lopez-del Burgo C. (2022), Matched and fully private? A new self-generated identification code for school-based cohort studies to increase perceived anonymity, “Educational and Psychological Measurement”, vol. 82 no. 3, pp. 465–481.

DiIorio 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”, vol. 23 no. 2, pp. 167–174.

Direnga J., Timmermann D., Lund J., Kautz C. (2016), Design and application of self-generated identification codes (SGICs) for matching longitudinal data, in: 44th SEFI Annual Conference, 12–15 September 2016, Tampere.

Grube J.W., Morgan M., Kearney K.A. (1989), Using self-generated identification codes to match questionnaires in panel studies of adolescent substance use, “Addictive Behaviors”, vol. 14 no. 2, pp. 159–171.

Kearney K.A., Hopkins R.H., Mauss A.L., Weisheit R.A. (1984), Self-generated identification codes for anonymous collection of longitudinal questionnaire data, “Public Opinion Quarterly”, vol. 48 no. 1B, pp. 370–378.

Krótkiewicz M., Wojtkiewicz K., Martins D. (2018), Influence power factor for user interface recommendation system, in: International conference on Computational Collective Intelligence, Springer International Publishing, Cham, pp. 228–237.

Kuzu M., Kantarcioglu M., Inan A., Bertino E., Durham E., Malin B. (2013), Efficient privacy-aware record integration, in: Proceedings of the 16th International Conference on Extending Database Technology, pp. 167–178, https://doi.org/10.1145/2452376.2452398.

Platje J., Palak R., Wojtkiewicz K. (2025), Sokrates forms. A research instrument for creating social impact of science on the example of system risk management, “The Central European Review of Economics and Management (CEREM)”, vol. 9 no. 1, pp. 47–62.

Sandnes F.E. (2021a), CANDIDATE. A tool for generating anonymous participant-linking IDs in multi-session studies, “PloS One”, vol. 16 no. 12, e0260569.

Sandnes F.E. (2021b), HIDE: short IDs for robust and anonymous linking of users across multiple sessions in small HCI experiments, in: Extended abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, Kitamura Y. (ed.), Association for Computing Machinery, Washington, pp. 1–6.

Vacek J., Gabrhelík R. (2024), Feasibility study of linking anonymous data of children in longitudinal school-based prevention research, “Addictology/Adiktologie”, vol. 24 no. 2, pp. 99–108.

Wiedenbeck S., Waters J., Birget J.C., Brodskiy A., Memon N. (2005), Authentication using graphical passwords. Effects of tolerance and image choice, in: Proceedings of the 2005 symposium on usable privacy and security, Association for Computing Machinery, Washington, pp. 1–12.

Wojtkiewicz K., Palak R., Telec Z., Litwinienko F. (2024), Modelling cognitive load of computer game users-case study, in: European conference on artificial intelligence, Springer Nature Switzerland, Cham, pp. 52–63.

Yurek L.A., Vasey J., Sullivan Havens D. (2008), The use of self-generated identification codes in longitudinal research, “Evaluation Review”, vol. 32 no. 5, pp. 435–452.