Political Analysis (), , 1–17

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Automated Survey Harmonization via Source-Consistent Entity Resolution

Dawson Verley,[*†] Amanda Kennard,[‡] Forrest Owens,[¶] Brandon de la Cuesta,[§] and Sara Constantino[|0]

†Energy and Resources Group, UC Berkeley, Berkeley, CA, USA
‡ Department of Political Science, Stanford University, Stanford, CA, USA
¶Department of Political Science, UC Davis, Davis, CA, USA
§Center on Food Security and the Environment, Stanford University, Stanford, CA, USA
|0Doerr School of Sustainability, Stanford University, Stanford, CA, USA
*Corresponding author. Email: verley@berkeley.edu

Abstract

Survey harmonization is a common challenge in applied social science research, particularly for multi-wave surveys in which question wording varies across rounds. Existing approaches rely on manual coding, making them difficult to scale and reproduce. We develop an automated procedure by framing survey harmonization as a problem of multi-source entity resolution. The approach combines sentence-embedding similarity with a source-constrained agglomerative clustering algorithm that enforces a one-per-source matching rule. Evaluated against a hand-coded crosswalk of 2,213 questions from seven Afrobarometer rounds, the method recovers the manual matches with a precision of 0.924 and a recall of 0.960 (F_1 = 0.941) at a small fraction of the cost of manual coding. Evaluated against the Latinobarómetro Corporation's own crosswalk, the method performs comparably, achieving a precision of 0.924 and recall of 0.994 (F_1 = 0.958). The procedure scales efficiently to thousands of items and applies more broadly to data integration problems involving multiple sources without reliable identifiers.

Keywords: Text as data; Survey harmonization; Entity resolution; Record linkage; Word embeddings; Sentence embeddings; Measurement; Cross-national surveys; Afrobarometer; Latinobarómetro

1. Introduction

In large-scale survey projects such as Afrobarometer and Latinobarómetro, many questions are repeated across survey rounds. These repeated items offer valuable opportunities to study change over time and to construct comparable measures across waves. However, combining data from such questions is difficult in practice. Survey questions are rarely assigned stable identifiers that persist across rounds, and even minor differences in punctuation, capitalization, or word choice prevent exact string matching. As a result, researchers cannot reliably use the question text itself as a unique identifier.

Existing approaches to survey harmonization become increasingly impractical as the number of rounds or survey locations grows. Although some projects provide harmonized datasets as a public good, their coverage is limited and updates are infrequent. For most applications, researchers must manually identify corresponding questions across datasets and construct their own crosswalks. While hand-matching may be feasible for a small number of items, it becomes intractable at scale. In major cross-national projects such as Afrobarometer, harmonization can require reviewing more than two million potential question pairs. This demands hundreds of expert person-hours, introduces substantial inter-coder reliability risks, and produces static outputs that are difficult to validate, document, and reproduce (Granda, Wolf, and Hadorn 2010).

In this paper, we develop an automated pipeline for multi-wave survey harmonization which we conceptualize as a special case of multi-source entity resolution. Our contribution is threefold. First, we formalize harmonization as recovery of a source-constrained partition over latent item-concepts, distinguishing an underlying item identity from its observed textual realizations across survey waves. Second, we introduce an end-to-end procedure that combines sentence-transformer embeddings (Reimers and Gurevych 2019) with a source-constrained agglomerative clustering algorithm. Third, we systematically evaluate the procedure using controlled simulations and labeled data from Afrobarometer and Latinobarómetro, benchmarking it against representative methods from three families of entity resolution algorithms, including probabilistic record linkage (Fellegi and Sunter 1969; Enamorado, Fifield, and Imai 2019), Bayesian multi-file models (Sadinle and Fienberg 2013), and graph-based clustering approaches.

Section 4 defines our target: a source-constrained partition of observed survey questions in which each cluster corresponds to an underlying item-concept. We operationalize this framework by representing each survey question with a pre-trained sentence-transformer embedding and computing cosine similarity between questions from different waves. Beginning with the most similar candidate pairs, the algorithm progressively merges questions into groups, provided that each group contains at most one question from any survey wave. This constraint allows the procedure to use information from all waves jointly: a potential pairwise match is accepted only when it is consistent with the other questions already assigned to the same group. The resulting matches are transitive, respect the known structure of the survey data, and can be computed efficiently. The complete Afrobarometer crosswalk can be generated in seconds on standard hardware.

We evaluate the procedure in three settings. First, Section 5 uses controlled embedding-space simulations in which the true partition is known and latent item-concepts are organized into batteries of semantically similar but distinct alternatives. Across the simulated conditions, our method achieves the strongest overall F_1 performance compared with four alternatives drawn from the three entity-resolution families described above. Second, Section 6 applies the method to labeled data from seven rounds of Afrobarometer, using an original manually coded crosswalk as the reference. The method achieves a precision of 0.924 and a recall of 0.960 (F_1 = 0.941). Third, we assess generalizability using eleven rounds of Latinobarómetro and an official crosswalk provided by the survey authors. The method achieves a precision of 0.924 and a recall of 0.994 (F_1 = 0.958), indicating that it transfers well across survey contexts.[1] Before presenting the method and its evaluation, we first describe the features of multi-wave survey data that make harmonization difficult and distinguish it from conventional record-linkage problems.

2. The Challenge of Multi-Wave Survey Harmonization

Survey harmonization refers to the process of determining which items across multiple waves of a survey–or across distinct surveys–correspond to one another or represent the "same" underlying item-concept. Granda and Blasczyk (2016) provides a straightforward definition of when two items should be considered to be the same: if they differ in *how* a concept is measured but not in *what* is measured. The strictest approach to survey harmonization might allow that only survey questions with identical wording and response scales be deemed to represent the same underlying item. Yet the exact wording of frequently asked items often evolves in minor ways over time such that they would fail this exact test despite measuring the same underlying item.

For example, survey organizations frequently carry questions forward across rounds while implementing minor changes to question wording or response scales based on feedback from enumerators or respondents. Survey researchers may substitute words for near-synonyms or restructure sentences by reordering clauses and shifting from active to passive voice. They may revise how the referent of a

1. Latinobarómetro results are based on English-language translations. The method achieves similar performance when applied to the original Spanish-language questions.

Table 1. Text of Eight Afrobarometer Questions. Response options are listed in the order they are coded; "don't know", "refused" and missing codes are omitted.

Round	Question No.	Question Text	Response Scale
1	dmpsay	In this country, you must be very careful of what you say and do with regards to politics.	Strongly agree; Agree; Neither agree nor disagree; Disagree; Strongly disagree
2	q41A	In this country, how often: Do people have to be careful of what they say about politics?	Never; Rarely; Often; Always
3	q53a	In this country, how often: Do people have to be careful of what they say about politics?	Never; Rarely; Often; Always
4	q46	In this country, how often: do people have to be careful of what they say about politics?	Never; Rarely; Often; Always
5	q56a	In your opinion, how often, in this country: do people have to be careful of what they say about politics?	Never; Rarely; Often; Always
6	q51a	In your opinion, how often, in this country: do people have to be careful of what they say about politics?	Never; Rarely; Often; Always
7	q42a	In your opinion, how often, in this country: Do people have to be careful of what they say about politics?	Never; Rarely; Often; Always
8	q40d	In your opinion, how often, in this country: Do people have to be careful of what they say about politics?	Never; Rarely; Often; Always

question is described, exchanging generic descriptors for specific ones, or shift from a conversational to a more formal register. Some questions may be compressed over time, while others are expanded to incorporate more precise instructions to the respondent. These changes alter the presentation, precision, or measurement scale of a question without substantially altering what the question is itself asking (Vila, Martî, and Rodrîguez 2011; Wolf et al. 2016; Fortier et al. 2017). In contrast, two questions which appear nearly identical in their wording yet differ in the subject of the question, the target population, or the time period to be considered can be said to measure fundamentally different things (Saris and Gallhofer 2014).

Consider the following example from the Afrobarometer project. Table 1 shows the text of eight questions, one from each survey round. Each question instantiates the same underlying item-concept: whether individuals must be careful when discussing politics. However, the questions have distinct variable names, are phrased in slightly different ways, and exhibit varying response scales. Whether to treat these questions as the same item in any analysis will depend on researcher judgment and the goals of the research project.

Matching questions across rounds where exact wording differs can be beneficial where it enables a researcher to expand the temporal or geospatial coverage of a question. The cost of doing so is measurement variance. Experimental research has long confirmed that even such minor alterations to item wording can shift the distribution of responses (Schuman and Presser 1981). By labeling two non-identical items as the same entity, a researcher implicitly accepts this variance in exchange for a

larger, more inclusive sample. The ideal definition of when two questions should be classified as the same then depends fundamentally on the purpose and goals of the research design. For the purposes of the empirical analyses below, we treat two questions as instances of the same item-concept when they ask respondents about the same substantive object, referent or target, population, and temporal frame, allowing differences in wording, syntax, punctuation, register, length, introductory language, and response format that do not alter the substantive interpretation of the item.

An advantage of our method is that it enables researchers to quickly and algorithmically implement their own desired level of stringency in the definition of sameness. This represents a significant advance over manual survey harmonization, the natural alternative to our algorithmic method. When manually harmonizing items across multiple waves or study locations, researchers must inspect question pairs one by one in order to determine whether they are sufficiently similar to be treated as the same underlying item. As the number of questions to be harmonized grows, this process becomes increasingly challenging. The example in Table 1 on its own is straightforward to adjudicate. Yet Rounds 2–8 of the Afrobarometer contain 2, 213 distinct questions. A researcher wishing to construct a comprehensive crosswalk must consider more than two million possible cross-round pairs, all the while ensuring that resulting matches are adjudicated consistently–possibly by multiple coders working in parallel–and without errors.[2]

We develop a procedure for multi-wave survey harmonization that turns a laborious, manual harmonization procedure into a semi-automated workflow that can run on a typical researcher's laptop in seconds.[3] Our method does not wholly eliminate the need for human judgment and validation. Nonetheless it dramatically improves the efficiency of harmonization while maintaining a high degree of flexibility, consistency, and accuracy.

3. Computational Approaches to Multi-Source Entity Resolution

Multi-source entity resolution is a challenging computational problem (we use the term "source" here to denote a single set of records, which in the survey context is analogous to a single wave). While efficient algorithms exist for pairwise matching between two sources, these have yet to be extended in ways suitable for multi-wave survey harmonization (Karp 1972; Spieksma 2000). The existing literature has developed three broad families of approaches, each with distinct trade-offs in terms of scalability, statistical interpretability, and suitability for text-based inputs.

The first and most widely used family is *probabilistic record linkage*, derived from the Fellegi–Sunter framework (Fellegi and Sunter 1969; Winkler 1988). These methods treat record pairs as draws from a mixture of matches and non-matches and estimate match probabilities based on comparison vectors constructed from multiple covariate fields. Extensions of this framework have been developed in political methodology (Enamorado, Fifield, and Imai 2019) and reviewed in recent surveys of the literature (Binette and Steorts 2022a). More recent work has adapted record-linkage methods to settings in which records contain only a single noisy or unstructured identifying field. Kaufman and Klevs (2022) develop an adaptive fuzzy string-matching approach for linking records using a single messy identifier while Ornstein (2026) incorporates pretrained text embeddings and cosine similarity into a probabilistic record-linkage procedure. These methods are designed primarily for pairwise linkage between datasets. Researchers wishing to harmonize large multi-wave surveys must reconcile matches jointly across many sources simultaneously all the while respecting transitivity and the underlying structure of the data.

2. As an empirical example of this challenge, in order to construct a benchmark Afrobarometer crosswalk used in Section 6 the authors employed four undergraduate research assistants for a total of approximately 500 hours, under the supervision of a postdoctoral researcher and a faculty member. Given the university's standard rates, the nominal cost of the research assistants' time in 2026 is more than $10,000. This does not take into account the cost of the postdoctoral researcher's or the faculty member's own time.

3. In the Afrobarometer, for example, our algorithm harmonizes 2,213 unique survey questions across seven waves in a quarter of a second.

A second family uses fully Bayesian generative models for multi-file entity resolution (Sadinle and Fienberg 2013; Aleshin-Guendel and Sadinle 2023; Marchant et al. 2021).[4] These models are well aligned with the structure of the problem as they natively accommodate multiple sources and provide principled uncertainty quantification. However, they face substantial computational challenges at scale. Posterior inference via Markov Chain Monte Carlo (MCMC) becomes impractical for datasets with thousands of records across many sources, particularly when compared to methods that operate in a single pass. Like probabilistic record linkage, these models also assume tabular inputs with multiple covariate fields rather than relying primarily on free-form text, making them poorly suited to typical large-scale harmonization tasks.

A third family frames entity resolution as a graph-clustering problem. In this approach, a similarity graph is constructed over candidate record pairs, and clustering or community-detection algorithms are used to partition the graph. Examples include simple thresholding of similarity scores, transitive closure, and community-detection methods such as Louvain (Blondel et al. 2008). These methods are computationally efficient and scale well to large datasets. However, they do not impose structural constraints linking records to their sources. As a result, they often produce overly large clusters by merging records that are jointly similar, leading to high recall (the true positive rate) but reduced precision (the proportion of correct matches among all returned matches).

What is missing from existing approaches is a method that combines scalability, compatibility with text-based similarity measures, and enforcement of the structural constraints inherent to multi-source data. The procedure we develop below addresses this gap.

4. A Source-Constrained Framework for Survey Harmonization

We formalize survey harmonization as recovery of a latent partition over survey items observed across multiple sources. Let N denote the total number of survey items, indexed by $i = 1, \ldots, N$. Each item i originates from one of K sources, such as survey waves or rounds, with source label $s_i \in \{1, \ldots, K\}$. Let q_i denote the *textual representation* of survey item i. We distinguish between an item's textual representation and the underlying *item-concept* that it instantiates, denoted by $c^{(i)}$.

We define item-concept identity using the substantive criteria developed in Section 2. An item-concept captures the features of a question that determine what is substantively being measured, rather than its particular textual realization in a given survey wave. The same item-concept may therefore have multiple textual representations across sources. Differences in wording, syntax, punctuation, register, length, introductory language, or response format may preserve item-concept identity when they do not alter the question's substantive interpretation. Conversely, questions with substantial lexical overlap may represent different item-concepts when they differ in their substantive object, referent, target population, or temporal frame.

This setup implies a straightforward generative structure. Survey design determines which item-concepts appear in each source and the textual representation of each. Let $\mathcal{C} = \{c_1, \ldots, c_M\}$ denote the set of distinct item-concepts represented in the data, and let $\mathcal{Q}(c_m)$ denote the set of textual representations associated with concept c_m. We assume that every observed textual representation corresponds to exactly one item-concept:

$$\text{for every } i \in \{1, \ldots, N\}, \qquad \exists!\, m \in \{1, \ldots, M\} \text{ such that } q_i \in \mathcal{Q}(c_m).$$

For the unique value of m identified above, $c^{(i)} = c_m$. Thus, identical textual representations imply the same item-concept, although the same item-concept need not have identical textual representations across sources. Table 1 provides a concrete example in which a single item-concept concerning freedom to discuss politics is instantiated through several different question texts across Afrobarometer rounds.

4. See Binette and Steorts (2022b) for a recent review.

Our estimand is the partition of observed survey items induced by their item-concepts. Define the block of observed item indices corresponding to item-concept c_m as

$$B_m = \left\{ i \in \{1, \ldots, N\} : q_i \in \mathcal{Q}(c_m) \right\}.$$

A block may be a singleton, corresponding to an item-concept that appears only once in the observed data. The target partition is

$$\Pi^* = \{B_1, \ldots, B_M\},$$

where

$$\bigcup_{m=1}^{M} B_m = \{1, \ldots, N\}, \qquad B_m \cap B_{m'} = \varnothing \quad \text{for } m \neq m'.$$

The first condition follows from the assumption that every observed representation belongs to an item-concept. The second follows from the assumption that no observed representation belongs to more than one item-concept. Together, these conditions ensure that every observed survey item belongs to exactly one block of Π^*. Because item identity is represented by a partition, it is necessarily transitive: if items i and j instantiate the same item-concept and items j and k instantiate the same item-concept, then items i and k must also instantiate the same item-concept.

4.1 The One-Per-Source Assumption

Throughout, we assume that an item-concept is represented no more than once within a survey wave. We refer to this as the *one-per-source assumption*. This corresponds to the clean-source or duplicate-free assumption used in some record-linkage settings (Zhang, Rubinstein, and Gemmell 2015). There are substantive reasons to expect it to provide a useful approximation in large, multi-wave surveys. Repeating the same item-concept typically increases survey costs without expanding the substantive scope of the resulting data (Groves et al. 2009). Repetition may also increase respondent fatigue and satisficing, potentially reducing survey quality and increasing nonresponse (Saris and Gallhofer 2014; Granda and Blasczyk 2016). Appendix A discusses the substantive realism of the assumption, circumstances in which intentional violations may occur, and possible adaptations of the method. Appendix B.5 evaluates the method's robustness to unintentional violations.

Let $s = (s_1, \ldots, s_N)$ denote the vector of source labels defined above. Formally, the one-per-source assumption restricts the set of admissible partitions to

$$\mathcal{P}^*(s) = \left\{ \Pi : \sum_{i \in B} \mathbb{1}\{s_i = k\} \leq 1 \text{ for every } B \in \Pi \text{ and } k \in \{1, \ldots, K\} \right\}. \tag{1}$$

Under this assumption, the target partition satisfies $\Pi^* \in \mathcal{P}^*(s)$. Our target is therefore a *source-constrained partition* of the observed survey items: two items originating in the same source cannot belong to the same block, even when their observed texts are highly similar.

4.2 Source-Constrained Greedy Agglomerative Clustering

For purposes of automated matching, we represent each observed text as a d-dimensional unit embedding, $x_i = g(q_i) \in \mathbb{S}^{d-1}$, where $g(\cdot)$ denotes the pretrained sentence-transformer model, so that $x_i^\top x_j$ is the cosine similarity between items i and j.

To identify the partition defined above, we adapt the greedy multipartite graph matching procedure of Zhang, Rubinstein, and Gemmell (2015) to a survey harmonization setting. The algorithm constructs a source-consistent partition by considering candidate cross-source links in

descending order of semantic similarity. Beginning with each survey item in its own singleton cluster, it iteratively merges the clusters associated with each candidate pair whenever doing so does not violate the one-per-source constraint. Candidate links that would produce a cluster containing more than one item from the same source are rejected. Only pairs whose similarity is at least λ are considered as candidate links.[5] An item with no such pair, or whose candidate links are all rejected by the constraint, remains in its own singleton cluster. The procedure leaves these singletons unmatched rather than forcing them into one or several groups.

The procedure is greedy in the sense that, when candidate links conflict under the source constraint, higher-similarity links receive priority over lower-similarity links. Computationally, cluster membership and the set of sources represented in each cluster can be maintained efficiently using a disjoint-set (Union-Find) data structure. The procedure is closely related to Kruskal's algorithm (Kruskal 1956), although here edges are processed to construct a constrained partition rather than a spanning tree.

Algorithm 1 Greedy Source-Consistent Entity Resolution

Require: Items $X = \{x_1, \ldots, x_N\}$ with source labels $s = \{s_1, \ldots, s_N\}$; similarity $U(x_i, x_j) = x_i^\top x_j$; threshold λ
Ensure: Partition $\Pi = \{B_1, \ldots, B_M\} \in \mathcal{P}^*(s)$
Compute $U_{ij} \leftarrow x_i^\top x_j$ for all $i < j$
Construct candidate edge set $E \leftarrow \{(i,j) : U_{ij} \geq \lambda \text{ and } s_i \neq s_j\}$
Sort E in descending order of similarity U_{ij}
Initialize Union-Find data structure with each item i in a singleton cluster $\{i\}$ with source set $S(\{i\}) = \{s_i\}$
for each edge $(i,j) \in E$ in sorted order **do**
    $B \leftarrow \textsc{Find}(i)$; $B' \leftarrow \textsc{Find}(j)$
    if $B \neq B'$ **and** $S(B) \cap S(B') = \emptyset$ **then**
        $\textsc{Union}(B, B')$
        Update source set $S(B \cup B') \leftarrow S(B) \cup S(B')$
    end if
end for
return Connected components of the Union-Find structure

4.3 A Worked Example

We provide a worked example of this algorithm for exposition purposes and to emphasize the importance of our one-per-source constraint. Figure 1 applies the procedure to six questions from three Afrobarometer rounds, all drawn from a single battery on perceived corruption. Write $q_{k,j}$ for the text of the jth question in round k, so that $s_i = k$ for that item. Most of these items have an unambiguous best match in each other round. But $q_{2,1}$—"elected leaders, such as parliamentarians or local councilors"—is almost equally similar to both round-4 items, and the round-4 item it matches best is not the one chosen by its own best round-3 partner. Linking every question to its most similar counterpart in each other round therefore places both round-4 items in the same cluster. Pairwise matching guarantees neither transitivity nor the one-per-source constraint. Sorting the same links by similarity and refusing any merge that would put two questions from one round together, as Algorithm 1 does, recovers the two series intact.

4.4 Manual Validation

Although the procedure thus far is fully automated, we recommend that researchers manually audit the output to verify accuracy. The algorithm itself provides a means of identifying clusters whose

5. The choice of λ is therefore substantive as well as technical: higher values favor precision by admitting only more similar candidate pairs, whereas lower values favor recall and broader coverage; Section 6.1 ("Selection of λ") examines this trade-off and provides practical guidance on threshold selection.

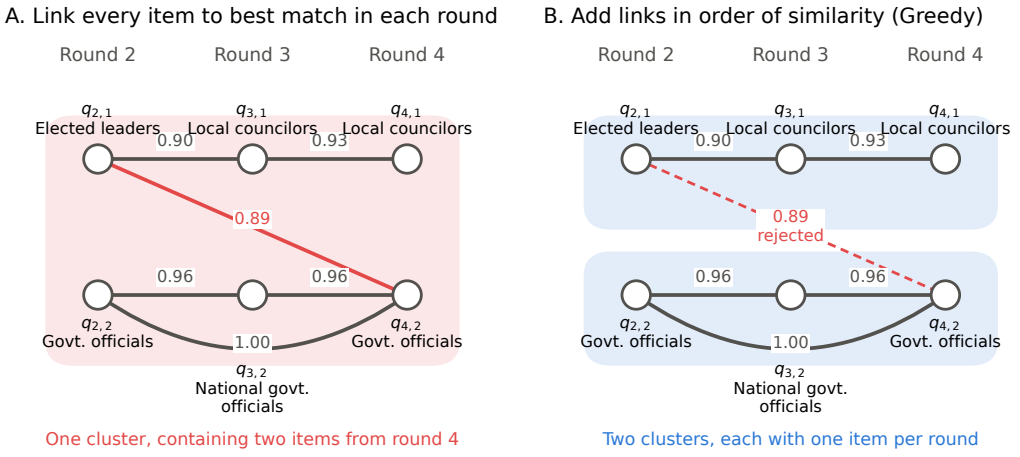

Figure 1. Pairwise matching violates the one-per-source constraint. Questions are drawn from the Afrobarometer battery "How many of the following people do you think are involved in corruption?" in rounds 2–4. Nodes are questions, arranged by round. Lines connect pairs of questions and are labeled with the cosine similarity of their sentence embeddings. Panel A links every question to its most similar question in each other round. The link between $q_{2,1}$ and $q_{4,2}$ (red) merges the two chains into a single cluster containing two questions from round 4, a violation of the one-per-source constraint. Panel B shows the result of the procedure above, which rejects the improper link.

membership is uncertain and which therefore are natural candidates for manual validation.[6] We describe this procedure in more detail in the context of our empirical results below.

4.5 Alternative Benchmarked Algorithms

In our empirical tests below we compare Algorithm 1 (Greedy) against four representative baseline methods drawn from the three families described in Section 3:

1. **Mutual Nearest-Neighbor Matching with Transitivity Correction (MNN-TC)**: For each pair of waves, retains a candidate link (i, j) with $x_i^\top x_j \geq \lambda$ only if j is the item in its wave most similar to i and i is the item in its wave most similar to j. Clusters are the connected components of the retained links. A post hoc transitivity correction (TC) then enforces the one-per-source constraint: within any component that contains two items from the same wave, links are removed in ascending order of similarity until no resulting cluster does.
2. **Thresholding with Transitivity Correction (Threshold-TC)**: Constructs a similarity-weighted graph $G_\lambda = (V, E_\lambda)$ with cross-wave edges $E_\lambda = \{(i, j) : x_i^\top x_j \geq \lambda, s_i \neq s_j\}$. Connected components are computed to form transitive clusters. The same TC step described above is then applied.
3. **Naïve Thresholding (Threshold)**: Constructs the same graph G_λ of cross-wave edges with $x_i^\top x_j \geq \lambda$ and returns its connected components. The resulting clusters are transitive, but the one-per-source constraint is not enforced.
4. **Louvain Community Detection (Louvain)**: Maximizes modularity Q on the weighted similarity graph G_λ using the Louvain algorithm (Blondel et al. 2008).

6. This approach borrows from Fellegi and Sunter (1969) which advises that pairs whose match probability falls between two thresholds are set aside for clerical review.

5. Simulation Evidence

Before turning to real survey data, we evaluate our algorithm against the four benchmarks described above within a setting where the true grouping of questions is known.[7] We adopt two separate approaches to the construction of simulation data. First, we generate artificial datasets in embedding space, a procedure which enables us to directly vary the difficulty of the matching problem and to examine how Greedy and the benchmark algorithms respond. Second, we generate artificial banks of survey item variations which we subsequently sample in order to again construct synthetic survey waves. The results of the latter simulation are described in Appendix C.

We construct simulated datasets containing 40 latent item-concepts observed across eight survey rounds. Each concept appears independently in each round with probability 0.8, conditional on appearing in at least two rounds, and generates at most one observed item per round. We organize the concepts into batteries of $B \in \{2, 4, 8\}$ semantically similar but distinct concepts, mimicking batteries of closely related survey questions. Two parameters govern the embedding geometry: $\mu_M \in \{0.75, 0.85, 0.95\}$, the expected cosine similarity between two observed representations of the same concept, and $\Delta \in \{0.30, 0.20, 0.10, 0.05\}$, the margin by which this exceeds the expected similarity between representations of different concepts in the same battery. Battery size provides an additional source of difficulty by varying the number of similar concepts competing within each semantic neighborhood. We generate 250 datasets for each of the resulting 36 design settings and evaluate every method using $\lambda = 0.75$. Appendix B provides the full data-generating procedure and additional results.

Figure 2 depicts the results. Greedy provides the strongest overall recovery, with its advantage widening as the similarity margin narrows and the number of competing concepts within each battery grows. In the setting closest to those observed in the Afrobarometer, with batteries of eight, Greedy achieves an F_1 of 0.98, compared with 0.93 for MNN-TC and 0.87 for Threshold-TC. Because all three methods have precision of at least 0.997, this advantage comes from recall: Greedy rejects only the inadmissible merge and continues considering weaker links, whereas the transitivity corrections remove links until the cluster becomes admissible, producing greater fragmentation. Methods that do not enforce the one-per-source constraint instead merge semantically similar concepts once their similarities cross the threshold and deteriorate sharply as batteries grow, approaching an F_1 of 0.2 with eight concepts. Finally, all methods lose accuracy when the margin becomes very small.

Appendix B, Figure 2 reports the same analyses using $F_{0.5}$, which places greater weight on precision than recall. Greedy remains the best-performing method in 31 of the 36 settings and is surpassed only by MNN-TC in five of the six most difficult conditions, which combine eight concepts per battery with a margin of $\Delta \leq 0.10$. Appendix B.5 shows that Greedy continues to perform well when the one-per-source rule is violated. When 5 or 10 percent of concepts are fielded twice in one round, its F_1 on the remaining concepts remains 0.98 and 0.97 respectively with batteries of eight. This is again higher than any of the benchmark methods.

6. Application to Existing Survey Data

We next evaluate our approach on two real-world survey harmonization tasks. First, we harmonize seven rounds of Afrobarometer data and compare the resulting matches with a manually coded reference crosswalk constructed by four undergraduate research assistants under the supervision of a senior researcher.[8] We use this crosswalk not as an error-free gold standard, but as the product of a realistic, resource-constrained manual harmonization process that researchers might otherwise employ. Second, we apply the method to eleven rounds of Latinobarómetro spanning two decades

7. Code for the simulation study and for all analyses in Section 6 and the Online Appendix was written with the assistance of Claude Opus 5.5 (Anthropic), a large language model. The authors specified each analysis and reviewed, tested and verified all code and results. The complete code is included in the replication materials. See Declaration on the Use of AI Tools below.

8. Constructing the crosswalk required approximately 500 hours of work and cost more than $10,000 USD.

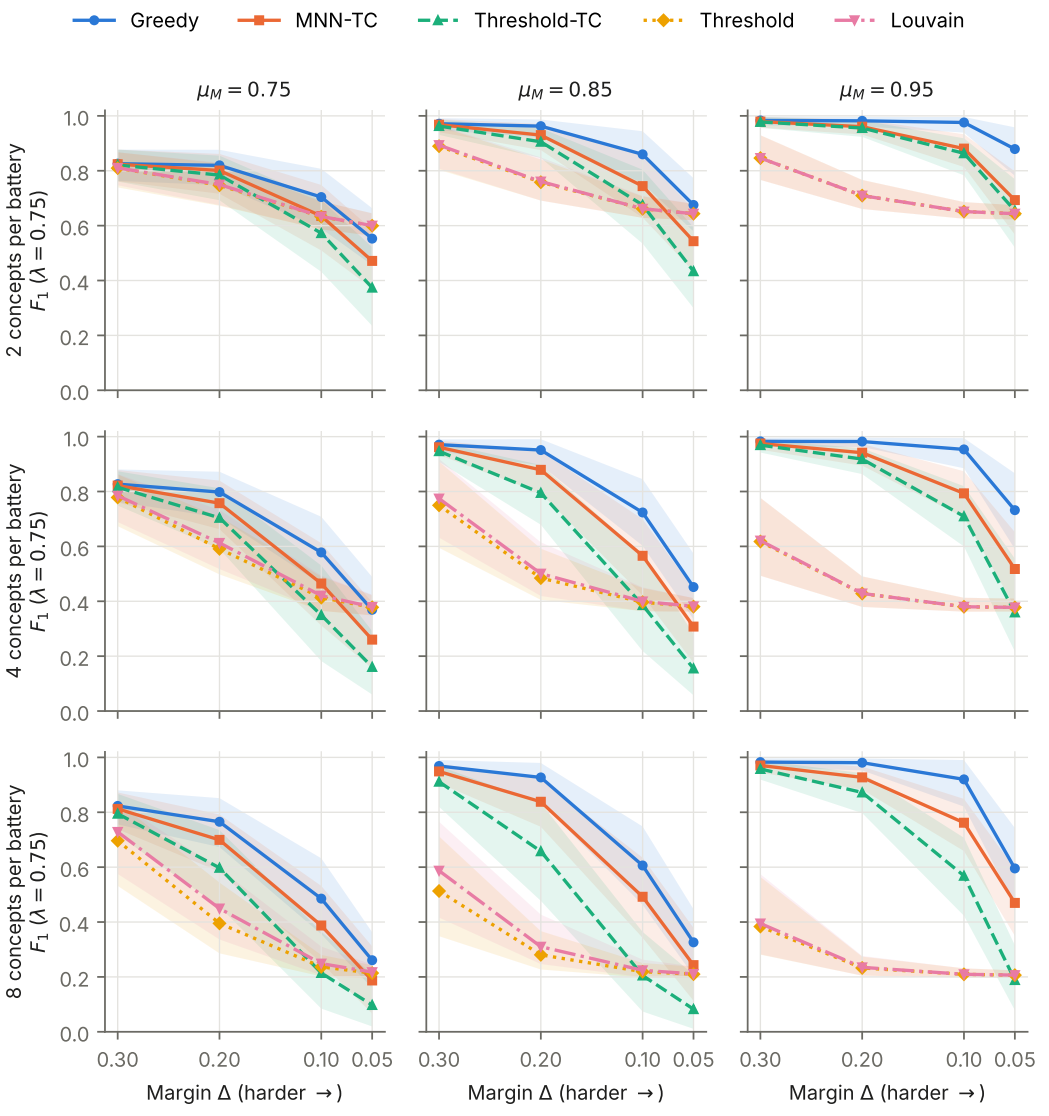

Figure 2. Recovery of a known partition in simulated embeddings. Rows: number of look-alike concepts per battery. Columns: average similarity between versions of the same question (μ_M). Horizontal axis: margin Δ between that average and the average similarity of look-alike concepts. Lines are means of F_1 over 250 simulated datasets and bands are 95% intervals across datasets. $\lambda = 0.75$ for every method.

and compare its output with an official crosswalk provided by the survey authors. Together, these applications assess performance against both a feasible manual alternative and an independently produced external reference. As the results below demonstrate, the procedure achieves high precision and recall in both settings.

Our first test uses data from the Afrobarometer project. We apply our method to the full set of questions from survey rounds 2–8. In our primary specification, we use the full text of each survey question, scraped directly from official codebooks. We also report results using only the short variable labels available in the survey data files. This comparison allows us to assess how sensitive performance

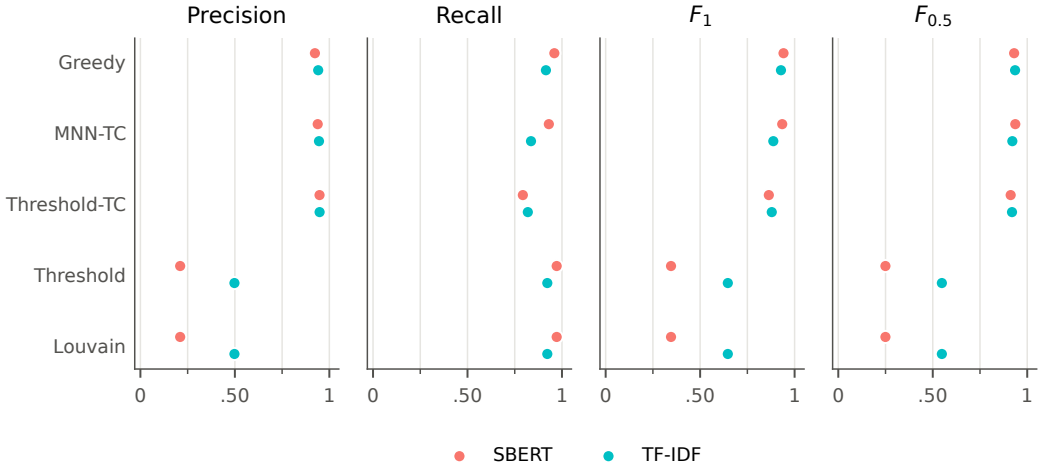

Figure 3. Precision, recall, F_1, and $F_{0.5}$ for five matching algorithms on the Afrobarometer labeled data at similarity threshold $\lambda = 0.75$, comparing SBERT (red) and TF-IDF (teal) similarity measures. Methods that enforce a one-per-source constraint (MNN-TC, Greedy, Threshold-TC) substantially outperform methods that do not (Threshold, Louvain). Within each method, the choice of similarity measure matters far less than the choice of matching algorithm.

is to the quality of textual inputs.

Table 2 presents key metrics describing the algorithm's performance while Figure 3 compares this to our four benchmarks. We implement all methods twice, using both Sentence-BERT and TF-IDF cosine similarity, setting $\lambda = 0.75$ throughout. We report precision, recall, F_1, and $F_{0.5}$. The last places twice as much weight on precision as recall and therefore favors methods that avoid false positives.[9] Using full question text, Greedy achieves a precision of 0.924 and recall of 0.960 ($F_1 = 0.941$; $F_{0.5} = 0.931$), the highest F_1 of any method tested. MNN-TC yields a slightly lower F_1 of 0.934 but the highest $F_{0.5}$ of 0.937, while Threshold-TC attains the highest precision (0.948) at the cost of approximately 17 percentage points of recall relative to Greedy. The broader comparison shows that algorithm choice matters substantially more than the choice between Sentence-BERT and TF-IDF: Threshold and Louvain preserve high recall but merge similar items into overly large clusters, producing low precision, whereas post hoc transitivity correction restores precision by discarding links and thereby reduces recall. Greedy provides the best overall balance. Performance remains strong but lower when only short variable labels are available ($F_1 = 0.890$; $F_{0.5} = 0.911$), and the complete comparison of 2,213 questions runs in approximately 0.25 seconds using `surveylinker`. Section 6.1 examines the sensitivity of results to our choice of λ.

To evaluate generalizability, we next apply the method to Latinobarómetro data. This dataset includes 2,490 questions across the first eleven survey rounds. We compare the resulting clusters to the crosswalk created by the Latinobarómetro Corporation. As with Afrobarometer, we rely primarily on the full question text for surveys through 2006.[10] The results are reported in Table 2. Using the English question text, the method achieves a precision of 0.924 and a recall of 0.994 ($F_1 = 0.958$, $F_{0.5} = 0.937$). Using the original Spanish text, it performs nearly as well ($F_1 = 0.945$, $F_{0.5} = 0.915$), suggesting that the approach does not depend on the availability of English-language translations. Implementation is again computationally efficient, completing in 116 seconds on a 2025 MacBook Pro with 32 GB of RAM. These findings suggest that the approach generalizes well across

9. Formally, $F_{0.5} = \frac{1.25 \times \text{Precision} \times \text{Recall}}{0.25 \times \text{Precision} + \text{Recall}}$.

10. Formatting changes to the codebooks after this year make it considerably more difficult to extend the analysis to contemporary waves.

survey contexts.

Finally, we perform a manual validation of our harmonized data, identifying a short list of clusters whose membership is least certain. The flagged set consists of clusters held together only by a weak link or adjacent to a strong candidate link rejected by the one-per-source constraint. On the Afrobarometer data at $\lambda = 0.75$, this procedure flags 109 of the 411 multi-item clusters. Appendix D provides details on the validation procedure and results. In a blind review, 25 percent of flagged clusters contained a question that did not belong, compared with just 3 percent of unflagged clusters. This contrast indicates that the diagnostic criteria successfully concentrate manual review on clusters most likely to require correction, improving the efficiency of *ex post* validation.

6.1 Selection of λ

Survey harmonization involves a trade-off between two kinds of error. A false positive groups items that measure different things and introduces measurement error into every analysis that uses the harmonized series. A false negative splits a cluster that should have been pooled, reducing the number of rounds, countries, and respondents available for the concept represented by a given question cluster. Which error is more costly depends on the research question. A study of measurement invariance, or of change in a narrowly defined attitude, should prioritize precision. A study whose goal is statistical power, or coverage of as many countries and years as possible, may reasonably accept some variation in question form in exchange for recall. In our procedure this trade-off is governed by the similarity threshold λ, which determines which cross-source pairs are considered candidate matches. Raising λ requires greater semantic similarity in order for a pair to be considered. This will generally increase precision at the expense of recall. Lowering λ allows less similar candidate pairs to be considered, increasing recall at the expense of precision.

As an example of this trade-off, consider the Afrobarometer's turnout question, asked in rounds 3–8. Rounds 6–7 add a preface noting that some people were unable to vote, and round 8 lists reasons for not voting; such prefaces are designed to reduce over-reporting (Belli et al. 1999; Duff et al. 2007). At $\lambda = 0.85$, Greedy pools only rounds 3–5 (about 72 country-rounds); at $\lambda = 0.65$, it pools all six (about 176). A study of turnout levels should keep the versions apart, while a study of its correlates may pool them. Appendix E.1 discusses this example in more detail.

Figure 4 explores the trade-off between precision and recall on the full Afrobarometer benchmark. Panel A shows that performance is flat over a wide range. F_1 stays between 0.93 and 0.95 for $\lambda \in [0.70, 0.85]$, and the threshold of 0.75 used in our main results was not tuned to the benchmark. $F_{0.5}$ in contrast is increasing at 0.75 and is maximized at $\lambda = 0.83$. Thus a precision-minded researcher might prefer to implement a higher threshold. Panel B plots the precision–recall frontier traced out by varying λ for the three methods that enforce the one-per-source constraint. For any precision target up to about 0.94, Greedy recovers more true matches than either alternative. At higher precision targets the three methods are close and trade places, with MNN-TC sometimes attaining

Table 2. Greedy (SBERT, $\lambda = 0.75$) on Afrobarometer and Latinobarómetro. Metrics are computed over unordered pairs of distinct questions. Afrobarometer: 2,213 questions from rounds 2–8, evaluated against the hand-coded crosswalk. Latinobarómetro: 2,490 questions from the 11 surveys fielded between 1995 and 2006, evaluated against the Latinobarómetro Corporation's crosswalk. [a]Spanish-language question text embedded with the multilingual model `paraphrase-multilingual-MiniLM-L12-v2`.

Survey	Input text	Precision	Recall	F_1	$F_{0.5}$
Afrobarometer	Full question text	0.924	0.960	0.941	0.931
Afrobarometer	Short variable labels	0.925	0.858	0.890	0.911
Latinobarómetro	Question text (English)	0.924	0.994	0.958	0.937
Latinobarómetro	Question text (Spanish)[a]	0.897	0.999	0.945	0.915

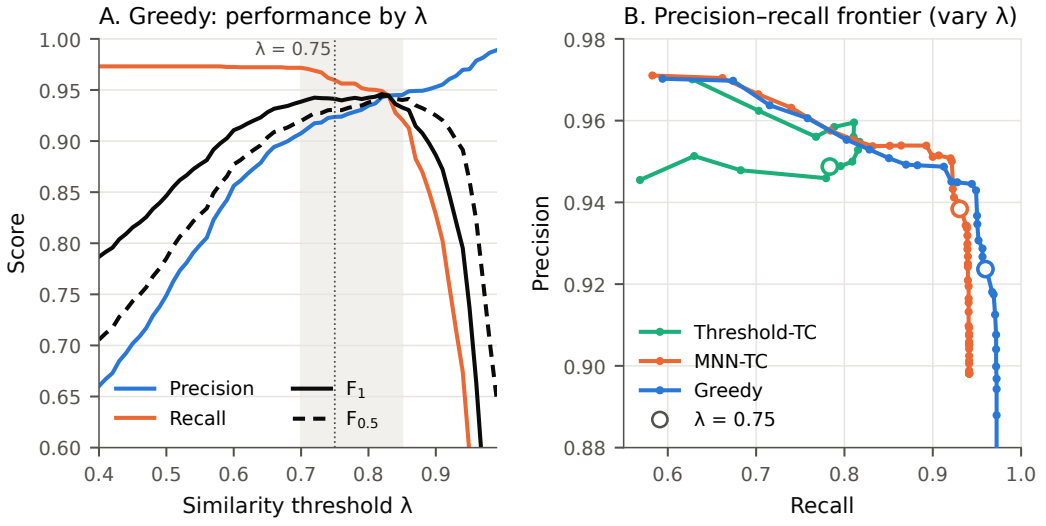

Figure 4. Sensitivity to the similarity threshold λ on the Afrobarometer benchmark (SBERT similarity, full question text). Metrics are computed over unordered pairs of distinct questions. Panel A: precision, recall, F_1, and $F_{0.5}$ for Greedy as λ varies from 0.40 to 0.99. The shaded band marks $\lambda \in [0.70, 0.85]$, over which F_1 varies by about 0.01. Panel B: precision–recall frontier traced out by varying λ in steps of 0.01 (0.02 for Threshold-TC) for the three methods that enforce the one-per-source constraint. Open circles mark $\lambda = 0.75$, the value used in the main results.

higher recall. A researcher who places overwhelming weight on precision can therefore raise λ, or choose among the constraint-enforcing methods, at a cost of recall and coverage.

How should researchers choose λ in a new application? Three considerations apply. First, 0.75 is a reasonable default for sentence-embedding cosine similarity.[11] On the Afrobarometer, performance is stable over a wide range in the neighborhood of 0.75. On Latinobarómetro, whose question wording changes little between surveys, performance continues to improve up to λ = 0.95 (F_1 of 0.977, compared with 0.958 at 0.75). Second, while it is not possible to optimally choose λ in cases where no ground-truth data exist, researchers should select a threshold based on how much wording varies within a survey series. This can be done subjectively according to researchers' familiarity with the survey product, or empirically by examining variation in question wording across the larger clusters after an initial run of the algorithm. Finally, because the matching step takes well under a second once similarities are computed, re-running the procedure across a grid of thresholds and inspecting which matches change is inexpensive, and we recommend it as standard practice.

6.2 Additional Analyses

In the appendix we report details on several further analyses. In Appendix E, we classify questions that (a) appear in only one round, (b) recur with identical wording, or (c) recur with changes in wording. The final category is where semantic matching provides its greatest advantage. In the Afrobarometer benchmark, 28 percent of cross-round links connect reworded rather than identical questions, of which Greedy recovers 88 percent. In Appendix F we investigate items on which Greedy and the benchmark disagree and provide a qualitative characterization of these items. In Appendix G, we provide a final demonstration of the empirical utility of the harmonized data by constructing and assessing twenty thematic public-opinion indices. These results illustrate how harmonization can be

11. For examples of this default setting in the literature see Weiner, Nashikkar, and Konar (2023) and Sentence Transformers (n.d.).

leveraged in order to construct measures at scale which vary across both time and space.

7. Discussion

This paper formalizes multi-wave survey harmonization as recovery of a source-constrained partition over latent item-concepts and introduces an automated procedure combining sentence embeddings with structure-aware clustering. Across controlled simulations and applications to Afrobarometer and Latinobarómetro, the method achieves high precision and recall while constructing cross-wave matches in seconds. It requires no task-specific labeled training data and can either produce complete crosswalks or identify candidate matches for targeted manual validation. We provide an open-source implementation in the Python package `surveylinker` to facilitate adoption and replication.

The method nevertheless has limitations. Its source constraint is appropriate only when each underlying entity appears at most once per source, and its deterministic assignments do not quantify match uncertainty. Future work should characterize its statistical properties, develop uncertainty-aware extensions, and study performance when the source constraint is imperfect. Despite this, our results demonstrate that combining semantic representations with substantively informed structural constraints can make large-scale survey harmonization substantially faster and more reproducible.

AI Disclosure

The authors used Claude (Opus 5.5), a large language model, to help write the code for all empirical analyses reported in this article and the Online Appendix.

Declaration on the Use of AI Tools

AI tools were used in the drafting of this manuscript as follows:

- **Code.** We used Claude Opus 5.5 (Anthropic; `claude-opus-5-5`), accessed through Anthropic's Claude applications (claude.ai and Claude Code) under a paid subscription, between August 1 and October 12, 2026 to write, debug, and document the code for the simulation study and for all empirical analyses reported in the manuscript and Online Appendix (see Sections 5 and 6).
- **Text.** We used ChatGPT (OpenAI; model GPT-5.6 Sol), accessed at chatgpt.com, between September 1 and October 9, 2026 to draft text in the Online Appendix and to proofread the manuscript, Online Appendix, and response memorandum.
- **Data.** The paraphrase banks used in Online Appendix C were generated with OpenAI's `gpt-5.6-terra` model via the OpenAI API in August 2026, as described in more detail within that section.

All code and text produced with these tools were reviewed, tested and revised by the authors, who take full responsibility for the content of this article. Neither tool was fine-tuned, trained on our data or otherwise modified; both were used through their standard interfaces. Both are proprietary models whose outputs are not exactly reproducible; the complete code and the generated paraphrase banks are therefore included in the replication materials. To our knowledge, no tool output reproduced third-party text, data or code requiring citation. The authors have no financial or other relationship with Anthropic or OpenAI.

References

Aleshin-Guendel, Serge, and Mauricio Sadinle. 2023. Multifile Partitioning for Record Linkage and Duplicate Detection [in en]. *Journal of the American Statistical Association* 118, no. 543 (July): 1786–1795. ISSN: 0162-1459, 1537-274X, accessed June 26, 2024. https://doi.org/10.1080/01621459.2021.2013242. https://www.tandfonline.com/doi/full/10.1080/01621459.2021.2013242.

Belli, Robert F., Michael W. Traugott, Margaret Young, and Katherine A. McGonagle. 1999. Reducing vote overreporting in surveys: social desirability, memory failure, and source monitoring. *Public Opinion Quarterly* 63 (1): 90–108. https://doi.org/10.1086/297704.

Binette, Olivier, and Rebecca C. Steorts. 2022a. (Almost) all of entity resolution. Publisher: American Association for the Advancement of Science, *Science Advances* 8, no. 12 (March): eabi8021. Accessed June 25, 2024. https://doi.org/10.1126/sciadv.abi8021. https://www.science.org/doi/10.1126/sciadv.abi8021.

———. 2022b. Modern Bayesian Entity Resolution [in en]. In *Wiley StatsRef: Statistics Reference Online,* 1–9. _Eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat08367. John Wiley & Sons, Ltd. ISBN: 978-1-118-44511-2, accessed June 26, 2024. https://doi.org/10.1002/9781118445112.stat08367. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08367.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, no. 10 (October): P10008. ISSN: 1742-5468. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Duff, Brian, Michael J. Hanmer, Won-Ho Park, and Ismail K. White. 2007. Good excuses: understanding who votes with an improved turnout question. *Public Opinion Quarterly* 71 (1): 67–90. https://doi.org/10.1093/poq/nfl045.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2019. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records [in en]. *American Political Science Review* 113, no. 2 (May): 353–371. ISSN: 0003-0554, 1537-5943, accessed June 25, 2024. https://doi.org/10.1017/S0003055418000783. https://www.cambridge.org/core/product/identifier/S0003055418000783/type/journal_article.

Fellegi, Ivan P., and Alan B. Sunter. 1969. A Theory for Record Linkage. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/do*Journal of the American Statistical Association* 64, no. 328 (December): 1183–1210. ISSN: 0162-1459, accessed June 25, 2024. https://doi.org/10.1080/01621459.1969.10501049. https://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049.

Fortier, Isabel, Parminder Raina, Edwin R. Van den Heuvel, Lauren E. Griffith, Camille Craig, Matilda Saliba, Dany Doiron, et al. 2017. Maelstrom research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology* 46 (1): 103–105. https://doi.org/10.1093/ije/dyw075.

Granda, Peter, and Emily Blasczyk. 2016. Data harmonization. In *Cross-cultural survey guidelines.* Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. https://ccsg.isr.umich.edu/chapters/data-harmonization/.

Granda, Peter, Christof Wolf, and Reto Hadorn. 2010. Harmonizing Survey Data. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts,* edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph. Mohler, Beth-Ellen Pennell, and Tom W. Smith, 315–332. Hoboken, NJ: Wiley.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology.* 2nd. Hoboken, NJ: John Wiley & Sons.

Karp, Richard M. 1972. Reducibility among Combinatorial Problems [in en]. In *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department,* edited by Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger, 85–103. Boston, MA: Springer US. ISBN: 978-1-4684-2001-2, accessed July 7, 2024. https://doi.org/10.1007/978-1-4684-2001-2_9. https://doi.org/10.1007/978-1-4684-2001-2_9.

Kaufman, Aaron R., and Aja Klevs. 2022. Adaptive fuzzy string matching: how to merge datasets with only one (messy) identifying field. *Political Analysis* 30 (4): 590–596. https://doi.org/10.1017/pan.2021.38.

Kruskal, Joseph B. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society* 7 (1): 48–50. https://doi.org/10.1090/S0002-9939-1956-0078686-7.

Marchant, Neil G., Andee Kaplan, Daniel N. Elazar, Benjamin I. P. Rubinstein, and Rebecca C. Steorts. 2021. D-blink: Distributed End-to-End Bayesian Entity Resolution. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2020.1825451, *Journal of Computational and Graphical Statistics* 30, no. 2 (February): 406–421. ISSN: 1061-8600, accessed June 26, 2024. https://doi.org/10.1080/10618600.2020.1825451. https://doi.org/10.1080/10618600.2020.1825451.

Ornstein, Joseph T. 2026. Probabilistic record linkage using pretrained text embeddings. *Political Analysis* 34 (2): 205–216. https://doi.org/10.1017/pan.2025.10016.

Reimers, Nils, and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 3982–3992. Hong Kong, China: Association for Computational Linguistics, November. https://doi.org/10.18653/v1/D19-1410. https://aclanthology.org/D19-1410.

Sadinle, Mauricio, and Stephen E. Fienberg. 2013. A Generalized Fellegi–Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2012.757231, *Journal of the American Statistical Association* 108, no. 502 (June): 385–397. ISSN: 0162-1459, accessed June 25, 2024. https://doi.org/10.1080/01621459.2012.757231. https://doi.org/10.1080/01621459.2012.757231.

Saris, Willem E., and Irmtraud N. Gallhofer. 2014. *Design, evaluation, and analysis of questionnaires for survey research.* 2nd ed. Wiley Series in Survey Methodology. Hoboken, NJ: John Wiley & Sons. ISBN: 978-1-118-63461-5. https://doi.org/10.1002/9781118634646.

Schuman, Howard, and Stanley Presser. 1981. *Questions and answers in attitude surveys: experiments on question form, wording, and context.* New York: Academic Press.

Sentence Transformers. n.d. *Retrieval.* https://sbert.net/docs/package_reference/util/retrieval.html. Accessed 2026-10-09.

Spieksma, Frits C. R. 2000. Multi Index Assignment Problems: Complexity, Approximation, Applications [in en]. In *Nonlinear Assignment Problems,* edited by Ding-Zhu Du, Panos M. Pardalos, Panos M. Pardalos, and Leonidas S. Pitsoulis, 7:1–12. Series Title: Combinatorial Optimization. Boston, MA: Springer US. ISBN: 978-1-4419-4841-0 978-1-4757-3155-2, accessed June 25, 2024. https://doi.org/10.1007/978-1-4757-3155-2_1. http://link.springer.com/10.1007/978-1-4757-3155-2_1.

Vila, Marta, M. Antônia Martî, and Horacio Rodrîguez. 2011. Paraphrase concept and typology: a linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural* 46:83–90. ISSN: 1135-5948.

Weiner, Bar, Aadi Nashikkar, and Soham Konar. 2023. *Optimizing multi-task classification finetuning in BERT: a multi-pronged approach.* CS224N Final Project Report. Stanford University. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/final-reports/final-report-169508330.pdf.

Winkler, William E. 1988. Using the EM algorithm for weight computation in the Fellegi–Sunter model of record linkage. In *Proceedings of the section on survey research methods,* 667–671. American Statistical Association.

Wolf, Christof, Silke L. Schneider, Dorothée Behr, and Dominique Joye. 2016. Harmonizing survey questions between cultures and over time. In *The SAGE handbook of survey methodology,* edited by Christof Wolf, Dominique Joye, Tom W. Smith, and Yang-chih Fu, 502–524. London: SAGE. https://doi.org/10.4135/9781473957893.n33.

Zhang, Duo, Benjamin I. P. Rubinstein, and Jim Gemmell. 2015. Principled Graph Matching Algorithms for Integrating Multiple Data Sources [in en]. *IEEE Transactions on Knowledge and Data Engineering* 27, no. 10 (October): 2784–2796. ISSN: 1041-4347, accessed June 25, 2024. https://doi.org/10.1109/TKDE.2015.2426714. http://ieeexplore.ieee.org/document/7095602/.