IJDC | *General Article*

Caring for Data's Soul – The Development of a Curation Impact Factor to Pinpoint the Effects of Data Curation Activities on Data Quality

Insa Bechert
GESIS Leibniz Institute for the Social Sciences

Kerstin Beck
GESIS Leibniz Institute for the Social Sciences

Ivet Solanes Ros
GESIS Leibniz Institute for the Social Sciences

Abstract

Curation matters for data quality! Hardly any survey data user would disagree with this statement. But how much of a difference it makes is difficult to count. In this paper, we will illustrate on the example of data from two cross-national social survey programs, the European Value Study (EVS) and the International Social Survey Programme (ISSP), the most common errors that occur in uncurated international comparative data and draw attention to the problems that can arise from such errors in analyses' results. To facilitate quality assessment and enable the assessment of data quality variation between countries within a survey, we developed a scheme that categorizes these errors, helps quantify them, and assigns them to possible curation measures. Based on this scheme, we developed an indicator that is called the Curation Impact Factor (CIF) that puts a concrete number on the data quality improvement due to curation effort and allows for comparability even across surveys. Therefore, the CIF could potentially be used to justify the use of resources for data curation in any survey data life cycle (e.g., in grant applications).

Submitted 12 August 2024 ~ *Accepted* 25 June 2025

Correspondence should be addressed to Insa Bechert. Email: insa.bechert@gesis.org

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The quality of scientific research results across all disciplines relies significantly on the quality of the data used in the research process. Survey data usually undergo a preparation process before they are made available to scientists. The people who organize this preparation process are called data curators. In Medieval Latin, the term "curatus" stands for a parish priest, someone "responsible for the care (of souls).[1]" Those who are inclined to poetry could say that, as data curators,[2] we are responsible for the data's soul. The question the analyses behind this article seek to answer is: How much improvement do we gain through our efforts?

This article is based on two social survey data examples coming from the International Social Survey Programme (ISSP)[3] and the European Values Study (EVS).[4] For both survey programs, data curation, integration of the national datasets, as well as member and user support, are carried out by the GESIS department Survey Data Curation (SDC)[5] in the International Studies team. The authors of this article are members of this team and have direct access to all internal data cleaning reports, which provide the basis for the process descriptions and error quantification.

During the course of an EVS or ISSP international survey data life cycle, a study wave is planned, data is collected within the countries, they are processed by the principal investigator's team (hereafter called the national team) according to the survey program's standards and, eventually, the national dataset, including documentation is deposited to GESIS SDC. Here, all national datasets are further processed, checked for errors, harmonized, integrated, documented, and released.[6] During all phases of data processing workflows, data preparation, checking, and correcting take place. This article concentrates on data preparation issues performed by data curators after the national teams have deposited their data, which means that the fieldwork institutes and national teams have finished their curation work. Although technical opportunities facilitate error detection and the execution of curation processes, data curators' human judgment is inevitable in program automatized routines.

Data curation is a broad term that has been defined in different ways by different actors (Knight, 2017). According to Johnston et al. (2014), data curation activities can be arranged around five steps of what they call a "data curation life cycle," which are: (1) Ingest; (2) Appraise (Accept); (3) Curate; (4) Access; and (5) Preserve. Our focus in this article is on the core curation tasks, usually performed by data curators at repositories, who are "caring" for the data by organizing, describing, cleaning, enhancing, and preserving them for public use (Inter-university Consortium for Political and Social Research (ICPSR), n.d.). This article aims to measure the success of the data curation process in terms of an increase in data quality and usability improvements.

The next section sheds some light on different aspects of data quality. It starts with an illustrative example of possible consequences if some errors are not detected before the data is published, and continues with some key aspects of the current discourse. The following section introduces the Data Curation Scheme. First, the Data, then the Curation, including checking procedures to discover different types of errors and the possible

---

[1] See the online Etymology dictionary at: https://www.etymonline.com/word/curate
[2] Also referred to as data managers or data processors.
[3] See the ISSP information webpages at: https://issp.org/ or https://www.gesis.org/issp
[4] See the EVS information webpages at: https://europeanvaluesstudy.eu/ and https://www.gesis.org/european-values-study
[5] In the case of the EVS the team is supported by data curators at the University of Tilburg (the Netherlands).
[6] Release and archiving processes are performed by the GESIS department of Data Services for the Social Sciences (DSS).

(re)action to those errors, and last, the Scheme. Finally, Curation Impact Factors (CIFs) are calculated to pinpoint the effect of data curation activities on data quality. Their advantages, shortcomings, and application areas are then discussed.

Data Quality

An Example of Error Consequences, or, in Other Words, "What If...?"

Some data errors have the potential to be more consequential than others. In the following section, we take the liberty to sketch a scenario of "what if the error had not been found and corrected before data publication."

Figure 1. An imaginable newspaper headline based on research on uncurated and uncorrected data.

Hypothetically, this could have been a newspaper headline created by a journalist using an uncurated EVS 2017 data file.

In Austria, as well as in Germany, national elections took place in 2017. In Austria, the Freiheitliche Partei Österreichs (FPÖ) achieved 26%, and in Germany, the Alternative für Deutschland (AfD) achieved 13%. Political scientists categorize both parties as right-wing extremists. Both parties stand for nationalistic and anti-immigrant policies (Grigat, 2017; Heinisch & Werner, 2019).

The following question was asked in the EVS:

Q51 Now we would like to know your opinion about the people from other countries who come to live in [your country] - the immigrants. How would you evaluate the impact of these people on the development of [your country]?

(v184)

5 – Very good
4 – Quite good
3 – Neither good, nor bad
2 – Quite bad
1 – Very bad

8 – don't know (spontaneous)
9 – no answer (spontaneous)

Figure 2. Original question Q51 (v184) taken from the EVS basic questionnaire.

The mean values for this item are 2.96 in Germany and 3.20 in Austria. This result assumes that the Austrian public is generally less xenophobic than the German public, which is a surprise, considering the comparatively high voting results for the nationalists in Austria. If you look at FPÖ-voters and AfD voters, respectively, the results are even more puzzling. The FPÖ voters reach a mean value of 3.84, which means that, on average, they almost reach a value that stands for "immigrants have a quite good impact on the Austrian society." However, German AfD voters reach a value of 1.84 between "quite bad" and "very bad," which is much more in line with what could be expected from voters of a right-wing extremist party.

What could be the reason for this, our newspaper journalist might wonder. Further research might lead them to an interview the Austrian political scientist Julia Partheymüller gave to the editor for politics, Ferdinand Otto, in the German online newspaper *Zeit Online* in 2019. She states that, although AfD and FPÖ are both scientifically classified as right-wing extremists, people in Austria were more resigned to it, that the party was generally seen as less radical, and that hardly anyone in Austria would ask anymore how right-wing the FPÖ is (Otto, 2019). Therefore, our journalist might conclude that maybe Ms Partheymüller is right, and the FPÖ and their voters are not that extreme and xenophobic after all, because the FPÖ has become a People's Party.

To get to the bottom of this matter, our journalist might even decide to conduct a regression analysis, checking for concrete correlations between voting for the FPÖ and attitudes toward immigrants, which would also confirm the results. There is a positive correlation between voting for the FPÖ and the view that immigration has a positive effect on the development of Austria.[7]

The real explanation is a simple error: the response scale was reversed by mistake in the Austrian field questionnaire used for data collection. Presumably, this has happened because values for this item were assigned in the opposite direction from most other questions in this questionnaire. The mistake was identified during the curation process after the data deposit, when test routines revealed suspicious correlations with other items. The same thing happened in five other countries. For the integrated data file, the data have been recoded according to the basic questionnaire, and the newspaper headline has been prevented.

What is Data Quality?

As the previous example shows, high-quality social science survey data is essential for scientists to conduct high-quality research, which, in turn, is essential for policymakers to make informed decisions that benefit society. However, what constitutes data quality? This question has been discussed intensely within and across research communities, stakeholders, and research funding agencies, and multiple dimensions have been identified (German Council for Scientific Information Infrastructures (RfII), 2020). For this brief excursion on the topic, we will take the perspective of data curators and concentrate on two main angles that have become evident in these discussions: (1) basic quality properties of the data itself; and (2) data fitness for research demands,[8] or as Wang and Strong (1996) put it, "intrinsic" and "contextual" data quality.

The key attributes of intrinsic data quality are accuracy and credibility, which basically means they should be error-free, complete, reliable, and valid. A framework that comprehensively covers possible errors in surveys is the Total Survey Error (TSE) framework, which has been developed (Groves, 1989), modified (Weisberg, 2005; Groves, 2009; Biemer et al., 2017), and supplemented (Smith, 2020) over decades. The scheme covers the potential errors creating bias, variance, or most commonly, a combination of both (Smith, 2011: pp. 466). The TSE literature is part of the extensive methodological literature in quantitative social research that deals almost exclusively with the

[7] For the regression results, see Figure A1 in the appendix.
[8] The ambiguity concerning the "fitness for research demands" approach is that "fitness" can be highly individual. What is meant here is the "fitness" for research demands of a respective research community and not the individual researcher.

quantification of reliability and validity of measurement instruments and data, while the accuracy of data is "taken for granted," although we know from practice that this simply cannot be assumed. While validity and reliability are not in the hands of data curators, many data errors can be detected by them through checking procedures. Even if only some of them can still be fixed, every correction improves the accuracy of the data. Data errors that cannot be corrected can at least be documented to alert researchers.

Contextual data quality refers to the suitability for a context-specific purpose (RfII, 2020, p. 16). For data journalists, timeliness and currency tend to be important. The relevancy of the data for the research topic is important for every kind of research. So is the accessibility, including findability and usability (Wang & Strong, 1996, p. 7). Essential for comparative research is the comparability of data across, for example, countries (Smith, 2011) or time (Chan, 1998; Davidov, 2008). Finally, a prerequisite for understanding the data and data quality is the precise documentation of data and metadata (RfII, 2020, p. 72).

Some of these data quality elements are not in the hands of the data curators. Timeliness only relies on them once the national data has been deposited. However, curators cannot always influence the time between data collection and the deposit from the national teams. In addition, comparability cannot be achieved retrospectively. If concepts are not comparable across countries, curators can only draw the attention of the responsible survey program committees to the issue when plausibility checks reveal suspicious data patterns. Moreover, most parts of the FAIR principles (Findable, Accessible, Interoperable, Re-usable), which are often mentioned to define data quality, rather point to the complex field of Research Data Management (RDM) and reach beyond the scope of responsibility of data curators as defined in this article. The most central area for data curators concerning contextual data quality is the transparent, accurate, and consistent documentation of data and metadata. Each piece of documentation helps researchers to analyze the data correctly and draw appropriate conclusions.

Beyond this, data curators for international survey data need to add another aspect to the definition of data quality, which is based on the standards set by the survey programs (usually the program's principal investigators, organized in working groups and in close collaboration with the team of curators). These standards define how the data is collected, which international coding schemes[9] might be used, and how the data should appear in the dataset. These comparability aspects are essential for defining data quality for international surveys.

To a certain degree, the research community communicates data quality to data curators through various channels, such as user conferences or when they get in touch asking for support. Generally, an active network characterized by close communication with users helps significantly in identifying the demands of the research community. Furthermore, conducting substantial research within the team of curators is essential to directly recognize user needs and possibly weaknesses in data quality. All the knowledge gained is fed into the preparatory phases and decision-making bodies of future surveys.

The next section introduces the survey data this article is based on, describes the curation processes established to improve the data quality aspects that are in the hands of data curators, and schematically presents a quantitative assessment of data quality and curation impact.

[9] Such as ISCO (International Labour Office (ILO), 2012) or ISCED (UNESCO Institute for Statistics, 2012).

The Data Curation Scheme

The Data

The EVS[10] and the ISSP[11] are large-scale, cross-national survey programs. While the EVS, as the name suggests, exclusively collects data in European countries, the ISSP is a global survey program. Both grew in membership from the 1980s, when they were founded, to 41 EVS and 45 ISSP member countries in 2024. Since their foundation, they have been collecting data on attitudes, values, and human behavior on topics important to the social sciences. The main difference is that the ISSP runs annual surveys on single topics, and the EVS conducts large surveys on several topics every nine years.

A fundamental difference concerning the handling of variables is rooted in the surveys' beginnings. While each EVS survey was planned as an independent survey, ISSP surveys were originally planned as comparably short questionnaires — the so-called modules — on specific topic areas to be attached to established national surveys.[12] To avoid asking respondents twice, background information (mostly demographic) was obtained from these national surveys. That means that only for the ISSP module variables was a clear coding frame agreed upon. Demographic information was collected according to the methodologies of different national surveys, and therefore, it was sometimes only roughly comparable with very heterogeneous documentation. During the expansion and further development of the ISSP, the background variables underwent several rounds of improvement and standardization. Nevertheless, ISSP countries are permitted to derive certain demographic information from country-specific variables; therefore, ISSP background variables still pose a special challenge for data curators.[13]

However, all surveys have their specific challenges. A challenge for data curators for the EVS 2017 was that four of the seven countries that collected data in both interviewer and self-administered modes (mixed mode) additionally used experimental matrix designs, in which the questionnaire was divided into several coherent blocks of questions. Respondents received some of these blocks in the first round of data collection and were usually asked to participate in a follow-up round to answer the remaining questions. Therefore, there are several cases in these surveys where no full interview was conducted (follow-up round declined, broken off, or not conducted at all). Although these "incomplete" cases from the matrix design surveys are only included in the EVS 2017 Matrix-Design Dataset and not in the EVS 2017 Integrated Dataset on which this paper is based, the data curation process, in general, turned out to be more complicated than usual.

To explore data curation activities, this article concentrates on one of each survey's datasets: ISSP 2013 – "National Identity III" (ISSP Research Group, 2015) with 386 variables and the pre-released EVS 2017 Integrated Dataset (EVS, 2020) with 466 variables.[14] Both datasets contain 34 participating countries.

---

[10] EVS website: https://europeanvaluesstudy.eu/about-evs/
[11] ISSP website at GESIS: https://www.gesis.org/en/issp/home
[12] The big national surveys of the founding countries were: Great Britain (British Social Attitude Survey (BSA)), USA (General Social Survey (GSS)), Germany (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS)), Austria (Sozialer Survey Österreich (SSOE)), and Australia (Australian National Social Science Survey (NSSS)).
[13] For further information on this issue, see the ISSP background variable history at: https://www.gesis.org/en/issp/home/issp-background-variables
[14] Documenting all data curation processes for further processing is a huge effort. This was performed in great detail for ISSP 2013. This preliminary work was used in this article. For the EVS, the current version was selected and the existing documents were further processed.

The Curation

Data processing at the GESIS SDC occurs in several phases.[15] The preparation phase starts with the implementation of upstream processes by developing documents of rules, goals, and processes that all players agree on before the survey is fielded.[16] For both survey programs, online portals, which play a central role in the preparation phase of a survey, have been set up. They serve as protected virtual workspaces for data and project management, communication management, and information transfer between the researchers involved in the project. They are a central location for tools and documents serving as prerequisites for ex ante harmonization. From the portals, so-called Standard Setups can be downloaded by the national teams. These Setups are generated within a statistical software program and display all variables, coding frames, and labels as they are supposed to appear in the final dataset. Portals serve as indispensable data quality tools because they provide the national teams with all the information they need and facilitate data and document exchange processes for all actors. However, the most important factor for a successful data processing workflow is close and timely communication between the national and curation teams.

Figure 3 shows an excerpt from the data curation phases.

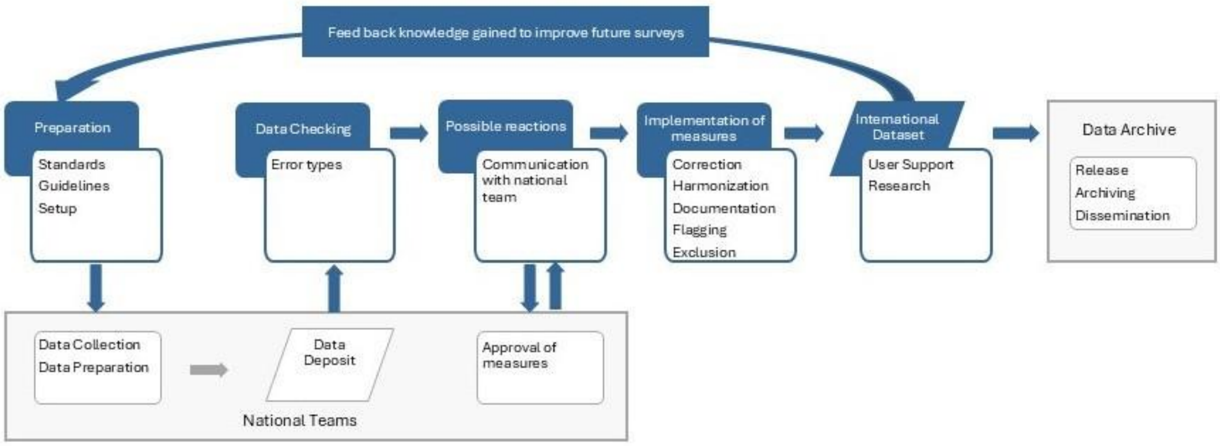

Figure 3. Curation phases. Phases indicated by grey boxes lie outside the core curation process.

After the data delivery, curators apply detailed test procedures to the data, communicate with national teams on all inconsistencies found, and implement all necessary changes. Finally, the single national datasets are combined and integrated into an international file. The amount of work in the data processing phase after delivery by the national teams varies depending on the experience and the time (resources) invested by the national teams. Therefore, it may well require several rounds of back-and-forth communication on data cleaning, during which the national teams review the data

[15] The following descriptions of curation processes are very survey-data-specific. It is not crucial to understand all these processes in detail to understand the subsequent calculation of the CIF and the related findings and conclusions.

[16] For the EVS, all documents elaborating the standards agreed on within the survey program are compiled in the *Methodological Guidelines*. One is the *Data Processing Guidelines* (GESIS, 2020, pp. 36), which clarifies all data processing steps according to EVS standards and defines the responsibilities of different actors. A basic document that sets standards for the ISSP background variables is the so-called *BV Guidelines* (ISSP Demographic Methods Group, 2024).

corrections made by the curators and approve their release.[17] All knowledge gained by curators during the curation process is incorporated into the preparation materials for future surveys and, if necessary, passed on to the relevant survey program committees to inform their basic decisions.

The following section describes the checking procedures applied. They take place on the four levels,[18] the Data Curation Scheme is based on:

- Case level;
- Variable level;
- Dataset level;
- Documentation level.

Checking procedures, shown here in more or less chronological order, vary slightly due to survey-program-specific rules and workflows. The aim is to check the compliance of national datasets with predefined survey standards and secure high data quality to ensure a reliable, integrated, international dataset.[19] This outline must be understood as a snapshot, because policies, as well as data processing practices, have changed and are changing over time.[20]

Checking procedures to discover different types of errors

When national survey data is deposited, the first check routine[21] concentrates on the completeness of the delivery. The data curator needs to check whether all mandatory data files, all required templates, and reports on data documentation have been deposited.[22] In addition, it is necessary to verify whether these files and documents are complete. That means, first, whether the data file contains all variables[23], and second, whether the documentation is up-to-date and comprehensive. If the delivery is complete, principal investigators from the national teams can be informed that the deposit was officially successful. A timely notification may be important because deadline compliance might impact certain rights or even the country's membership in the survey program.

After verifying the completeness, curators examine the frequencies of all variables contained in the national datasets. At first glance, it becomes obvious whether any variable

---

[17] That was a long and cumbersome process described in a nutshell. For a more comprehensive overview of Data Curation procedures, including models of Data Curation Life Cycles see Higgins (2008).

[18] See Table 1, Column 1.

[19] An example for survey-specific data processing policies regarding data cleaning: while the ISSP agreed on the rule that data should not be cleaned completely and individuals are allowed to answer inconsistently, the EVS set the rule stipulating that, after an investigation by the national teams, data should be only corrected if less than 1% of cases are affected by inconsistency in a sample. If more than 1% of cases are affected, and a systematic deviation can be assumed, they should not be cleaned but documented.

[20] Inconsistencies and errors found are categorized in Table 1, Column 2.

[21] Data checks are mainly conducted using the statistical software program SPSS. Some syntax jobs require STATA or R. Beyond; Excel is used for the documentation of check results. Although technical opportunities facilitate error detection and the execution of curation processes, data curators' human judgment is inevitable in programming automated routines.

[22] An example: a complete ISSP deposit comprises one dataset, one output by a statistical program, and four text documents that include a method report, a characterization of national characteristics as a reference for the sample, a description and translation of the background variable measures, and the original national questionnaire.

[23] Missing variables occur more often in the ISSP than in EVS, because the collection of demographic variables is regulated more strictly in the EVS, and the questions on demography are part of the basic questionnaire.

or value labels are wrong or even missing. The frequencies are controlled to discover any conspicuous particularities. In the next curation phase, data are controlled for methodological and/or technical errors, such as incomplete cases[24] or duplicated IDs. Duplicated IDs occur when either an ID number was issued twice or the whole case, including the ID number, appears twice in the data. In the latter case, we speak of duplicate records that can occur due to falsification or, equally likely, a technical mistake. Another check routine was specifically developed for testing the data for so-called "near duplicates"[25], which identifies cases that are not identical but more similar to each other than is statistically likely and, therefore, raises the suspicion of coding errors or data fabrications.

Another set of check routines deals with data plausibility within national datasets. Variables between which a theoretically based direction of the relationship can be assumed and has been confirmed by earlier data should be correlated with each other.[26] If not, further investigation is warranted. A similar test is applied for checking the direction of long scales (usually 11 points). It is a common mistake that such scales are reversed, either in the design of the national questionnaire or during the coding process. For the identification of these errors, international survey programs can make use of cross-country data comparisons. If correlations in single countries do not show the expected direction or strength, it is reasonable to assume that a reversed scale error occurred. For background variables, which are similarly collected in each wave, the respective check-syntax-jobs are well established, and the test can be automated. For substantial module variables, plausibility test routines must be set up for each wave.[27]

Plausibility tests may find impossible values or combinations of values. In addition, they might reveal so-called outliers. Outliers are values that appear highly unlikely but could theoretically be true, such as respondents who report having 18 children or being widowed at the age of 19. It must be evaluated carefully whether these are true values or not. These tests also contain a check for so-called wild codes. These are codes appearing in the data that simply do not exist in a predefined scheme. Wild codes are most likely to appear when comprehensive (international) coding schemes, such as the ILO/ISCO scheme for occupational classifications (ILO, 2012) are applied to data. They are usually coded in by well-meaning coders in the countries who cannot find a proper code in the scheme for a certain country-specific occupation, and add a new code that fits in their opinion.

Another common source for potential survey errors that needs to be checked carefully is filter routings. Filter routing errors occur when interviewers or, in self-administered interviews, respondents overlook the filtering instruction of a question for follow-up questions[28] and, consequently, answer incorrectly or do not answer them, although they were supposed to do so. The reconstruction of the missing values according to the routing rules is necessary to perform other check procedures, such as correctly calculating the completeness of interviews.

Data curators at GESIS SDC strive to identify errors, track down their sources, and provide solutions by recommending data recodes or documentation of the deviations. If the solution involves changes to the data, it is situational who makes those changes. Sometimes, the principal investigators send an updated dataset after a request. Sometimes,

[24] The definitions of how many missing answers in an interview make an incomplete case vary. ISSP and EVS orient towards benchmarks set by the American Association for Public Opinion Research (AAPOR); see AAPOR (2023).

[25] The calculations are based on Kuriakose and Robbins' (2016) similarity assessment algorithm, and the "Percentmatch Approach" was conducted with STATA. Based on this approach, the match percentage for a case from any possible pair of cases within a sample is given by the number of identical values across the pair, divided by the number of variables in the comparison.

[26] For example, family income should have a positive correlation with social status, and people who are retired should be older than those in the labor force.

[27] An example from the ISSP module on religion: the individual level of religiosity should correlate positively with church attendance.

[28] Or when the filtering function is mis-implemented in the computer-assisted questionnaires.

it is more expedient if the curator does the actual recoding after the principal investigator has signed off on the measure. The range of possible errors is, of course, wider, and new, unexpected discrepancies are constantly emerging, some of which can only be discovered through a sound instinct for the data.

The following section describes the possible (re)actions of curators to the inconsistencies and errors found during the checking process.

Possible (re)actions to errors

The multifaceted nature of errors and deviations requires a wide range of case-specific courses of action.[29] If an interview has too few valid responses to be considered complete, the case must be deleted or at least flagged, so that researchers may decide for themselves whether to use them or not. Clearly identified duplicates or "near duplicates"[30] must be deleted. Routing errors are tricky. Superfluous information can be deleted to maintain the comparability of data structures between countries. Information not given by mistake, however, cannot be generated subsequently, and therefore, it must be documented that parts of the sample were erroneously excluded from the question. In general, curators aim to keep as much information as possible and allow respondents to be inconsistent in their response behavior. If plausibility tests uncover obvious errors, data should be recoded in case the true value is known. If not, which is usually the case, the only options are the deletion of information or documentation. If the test reveals peculiarities, it is up to the national teams to investigate the source and decide whether any action is warranted.[31]

Variable names, variable labels, value labels, and variable format definitions that do not follow the international standard, as well as deviating formats and missing definitions, must be adjusted by recoding. The decisions for actions become less obvious if a country uses response categories that deviate from the predefined standard. If a category is missing or there are too many options for responses, the deviation must be documented. One option to deal with such data is to offer the variable as a country-specific variable; the second option is to label surplus categories as country-specific within the variable. The third option is to recode the deviant coding scheme so that it matches the standard with harmonization techniques, such as linear stretching (de Jonge et al., 2014).[32]

In cases involving wild codes and outliers, decisions depend significantly on the availability of country-specific expertise. Without such expertise, it is neither possible for data curators to appropriately reassign, for example, an invented country-specific occupation code to a valid code in a predefined scheme, nor is it possible to judge whether, in specific survey constellations, unlikely values might be rational.

Decisions at the dataset level are usually easier to make. If a variable is missing (e.g., by a processing mistake before the deposit), that means the information was collected, and the variable can be submitted subsequently by the national teams. If the information was not collected, the lack of data can only be documented. Sometimes, missing variables can be generated based on related variables, such as computing the age of the respondent using the date of birth and the date of interview. Variables that are missing in the national dataset need to be created and coded to the appropriate missing value for the international file.

Finally, data curators deal intensively with the supplied documentation. Any inconsistencies with the data must be investigated and corrected, because only correct

---

[29] Column 3 of Table 1 assigns the possible actions to the potential errors. For a comprehensive overview of checking and handling data errors, see Brislinger and Moschner (2019).

[30] The duplicate and near-duplicate tests were implemented in the ISSP check routines in 2017. All datasets were checked retrospectively, and suspicious cases and recommendations for actions were documented. The Overview_Duplicated_Records Document can be downloaded with each affected data file. In updated ISSP datasets, duplicate records are deleted.

[31] For a comprehensive overview of plausibility checks and measures in survey data, see Bechert et al. (2023).

[32] Here, it needs to be considered that response behavior differs according to the number of categories given. This means that the data can only be compared with reservations.

information on survey methods, question texts, measurements, and variables makes responsible handling of the data possible.[33]

Data processing and data cleaning documentation procedures differ across surveys. Therefore, comparisons of curation measures based on data quality problems that are detected by applying the check routines are not that easy. The following section, nevertheless, attempts to do just that by recording the number of actions undertaken during the curation processes of ISSP 2013 and EVS 2017 in the scheme presented in Table 1.

The Scheme

The Data Curation Scheme presented in Table 1 introduces a framework that combines the errors and measures described in the previous sections[34] and quantifies the frequencies with which curative interventions were necessary during the data preparation processes. This quantification is a difficult process. Some errors are common and easy to categorize; however, others are multilayered and might fall into different categories at the same time. In addition, the consequential reactions can overlap. For the errors, individual decisions have been made to categorize them as precisely as possible and as similarly as possible across survey programs. The framework for curative means covers all potential reactions that might be used slightly differently in ISSP and EVS contexts. Although the quantifications produce a rather rough measure, they shed some light on how many potential quality problems have been identified and removed during the data curation processes.

Column 1 “Level” specifies the four levels on which errors occur: case level, variable level, dataset level, and documentation level. On the case level, check routines concentrate on completeness and case consistency. On the variable level, the focus is on not prespecified or missing values, but on the dataset level, checks are related to the structure and completeness of data files. Errors, in this context, are all deviations from the standards set by the survey programs. At the documentation level, checks focus on the accuracy of metadata and the consistency of data and documentation.

Column 2 “Type of error” indicates the specific type of error detected by the checking procedures explained above.

Column 3 “Type of possible (re)action” assigns the curators’ (re)actions of either correcting the error, if possible, or, if this is not possible, of making data users aware of them through documentation.[35] Of course, all actions that include data alternation, which are not simple corrections but could be of interest to data users, are also being documented. However, the table captures the action of documentation only if this is the only possible action.

Columns 4 “ISSP” and 6 “EVS” assign the concrete numbers to these actions. If a cell shows “n.a.” (not available), there was no checking routine for this error in this year for this survey program. Columns 5 and 7 present Herfindahl Index (HI) values,[36] a concentration measure, for the distribution of necessary curative actions across the different country samples within each survey dataset. We rescaled the measure to range between 0 (low concentration) and 1 (high concentration). Because the versions of the datasets we analyzed both contain 34 countries, the index values are comparable.

---

[33] In addition, see RfII (2020, p. 72), stressing the importance of documentation.
[34] The list covers procedures after the incoming checks for completeness have taken place.
[35] SDC survey documentation uses codebooks, called Variables Reports (PDF), and online documentation via: https://search.gesis.org
[36] The HI is a concentration measure developed by the economists Hirschman and Herfindahl in 1945 and 1950. Ranging from 0 to 1, it is applied in multiple contexts, for example, when calculating a firm’s market share to capture the degree of the market's tendency toward monopoly (maximum) or perfect competition (minimum), or in measuring the concentration of wealth in households (Rhoades, 1993).

Table 1. Potential errors in survey data, types of actions, and quantification of ISSP 2013 and EVS 2017 (34 countries in each survey).

Level	Type of error	Type of possible (re)action	ISSP	HI	EVS	HI
Case level	Incomplete interview	Flag	n.a.	–	223	0.04
		Exclude				
	Duplicate or near duplicate	Document suspicious cases	3	0.73	–	1
		Exclude suspicious cases	10		1	
	Missing/multiplied respondent ID	Recode all ID numbers but one	–	–	2	0.73
Variable level	Routing error	Recode	137	0.01	455	0.01
		Document	–		88	
	Wrong derivations	Recode	81	0.08	52	0.04
		Document	–		–	
	Deviations from the international standard, including wild codes: variable names and labels, value labels	Recode and/or relabel	894	0.07	297	0.03
		Document	–		22	
		Compute flag variable	–		98	
	Plausibility inconsistencies, including outliers	Recode	–	0.12	150	0.01
		Document	2		179	
		Exclude	–		–	
	Missing / or erroneously added response categories	Compute national-specific variable	8	1	76	0.08
		Recode categories	–		25	
		Document	–		–	
	Incorrect missing definitions	Recode and/or relabel	159	0.19	835	0.03
		Document	–		–	
		Compute flag variable	–		42	

	Non-achievement of measurement goal	Compute flag variable	–	0.31	6	0.48
		Document	38		–	
Dataset level	Incomplete set of variables	Add	67	0.04	20	0.19
		Create variable	14		44	
	Order of variables	Rearrange	–	–	–	–
		Document	–		–	
Documentation level	Inconsistency: data and documentation	Correct	30	0.06	140	0.02
	Incorrect metadata	Correct	4	0.36	39	0.02
		Document	–		7	
SUM			1,457		2,801	

The numbers in Columns 4 and 6 show that some errors are more common than others. On the case level, routing errors are by far the most frequent type of error in both surveys. On the variable level, deviations from the international standard and incorrect missing definitions are the most common errors.[37] In addition, plausibility inconsistencies occur frequently, especially in the EVS.[38] But also in the ISSP, 32 cases of plausibility inconsistency have been detected. After discussions with the national teams, however, 30 of these cases remained untreated because there was no evidence that the entries were definitely incorrect. At the dataset level, incomplete national datasets were submitted in both surveys. The level of incompleteness "by mistake," implying that the missing variables could be delivered later, is much higher in the ISSP. The number of inconsistencies discovered between the data and documentation was not low in either survey. All could be corrected. Our revised scale error from the previous newspaper headline example was detected during plausibility checks and appears among the 150 EVS recodes in the scheme.

The HI values give us an idea about the distribution of errors across countries, indicating that they are not equally distributed. In other words, the data quality differs across countries. Some errors happen almost everywhere; for example, routing errors, deviations from the international standard, or inconsistencies between data and documentation have low index values in both surveys. Other errors are rare and only occur in very few countries. For example, 13 duplicate cases have been discovered in the ISSP data, but only in two countries, which generates an HI of 0.73. The one duplicate in the EVS or the eight ISSP variables with missing categories occurred in only one country. Consequently, the HI has a value of 1.

The Curation Scheme captures the number of actions required per country.[39] Based on the scheme, we can easily see the curation effort taken and estimate the curation effort for future survey waves. The sum of actions taken for the ISSP was 1,487, and for the EVS it was 2,801. What does the sum of errors teach us? Is the data quality of ISSP 2013 almost twice as good as that of EVS 2017? Probably not. On the one hand, there may be survey-specific strategies. For example, the threshold at which an implausibility is documented may be lower in the EVS than in the ISSP. On the other hand, and that affects the comparison even more, the quantified effort is comparable across countries within a survey but not across surveys because of different numbers of variables and cases. To achieve comparability, in the next section, the quantified curation effort of both surveys will be put into proportion.

Developing a Curation Impact Factor

The previous sections illustrate the huge effort required in data curation. However, it takes more than a snapshot to pinpoint the effects of data curation activities on data quality. Therefore, a CIF was developed to capture the quality difference between uncurated and curated data. This factor calculates the percentage of improvement for each

[37] Parts of the outstandingly high number of incorrect missing definitions in the EVS originate in System Missings that had to be recoded mostly to "Not applicable."

[38] The high figures here are mainly due to error-prone measures of the household composition.

[39] The scheme concentrates on the most common errors and curative (re)actions. Exceptional issues that occur in every survey and can be extremely time-consuming, precisely because they do not happen regularly, cannot be covered by the scheme. In addition, tasks that are peripheral to the actual curation work, such as teaching students in the national teams elementary data processing procedures, are not covered here. Finally, not covered, is the fact that a dataset is never really finished. If remaining errors are discovered, an erratum must be written and published, and the data must be updated eventually.

curation level. Their mean value indicates the average percentage of improvement for the complete international dataset.

To calculate the CIF, the possible maximum value for each curation measure (either the total number of cases per survey, the total number of variables per survey, or the total number of required documents per survey) is taken and put the actual number of curation measures taken (indicated in the Curation Scheme in Table 1) in relation to this to obtain a percent value.

On the **case level,** that means for (1) ISSP 2013 and (2) EVS 2017:

$$(0)\ \ CIF = \frac{Number\ of\ curation\ measures\ \times 100}{Number\ of\ cases}$$

$$(1)\ \ CIF = \frac{13 \times 100}{46{,}935} = 0.03 \qquad (2)\ \ CIF = \frac{226 \times 100}{56{,}506} = 0.40$$

For the ISSP (1), in the curation scheme category “case level,” we acted on 0.03% of cases to improve the data quality. For the EVS (2), we acted on 0.4%.

On the **variable level** (variables per country), that means for (1) ISSP 2013 and (2) EVS 2017:

$$(0)\ \ CIF = \frac{Number\ of\ curation\ measures\ \times 100}{Number\ of\ variables}$$

$$(1)\ \ CIF = \frac{1{,}319 \times 100}{3{,}740} = 35.27 \qquad (2)\ \ CIF = \frac{2{,}325 \times 100}{10{,}574} = 21.99$$

For the ISSP (1), in the curation scheme category “variable level,” we acted on 35.27% of variables to improve the data quality. For the EVS (2), we acted on 21.99%.

The calculation of the CIF for the **dataset level** diverges from the other levels because, despite the errors concerning the dataset structure, we count the variables that we had to act on. Errors that had been counted on a dataset level did not occur in these two survey waves. That means for (1) ISSP 2013 and (2) EVS 2017:

$$(0)\ \ CIF = \frac{Number\ of\ curation\ measures\ \times 100}{Number\ of\ variables}$$

$$(1)\ \ CIF = \frac{81 \times 100}{3{,}740} = 2.17 \qquad (2)\ \ CIF = \frac{64 \times 100}{10{,}574} = 0.61$$

To improve the data on the dataset level, we acted on 2.17% of the variables in the ISSP dataset and 0.61% of the variables in the EVS dataset.

On the **documentation level,** that means for (1) ISSP 2013 and (2) EVS 2017:

$$(0)\ \ CIF = \frac{Number\ of\ curation\ measures\ \times 100}{Number\ of\ required\ documents}$$

$$(1)\ \ CIF = \frac{34 \times 100}{102} = 33.33 \qquad (2)\ \ CIF = \frac{186 \times 100}{309} = 60.19$$

To improve the documentation quality, we acted on 33.33% of the required documents deposited for ISSP 2013 and 60.19% of the documents deposited for EVS 2017.

Table 2. The CIF

Levels	Improvement in % ISSP 2013	Improvement in % EVS 2017
Case level	0.03	0.40
Variable level	35.27	21.99
Dataset level	2.17	0.61
Documentation level	33.33	60.19
CIF – total (sum/4)	17.70	20.80

The calculations show that errors have been reduced, and data quality has been improved for ISSP 2013 by 18% and for EVS 2017 by 21%.

The next section, on the one hand, discusses the CIF's shortcomings, which need to be considered when interpreting the results above. On the other hand, it discusses the value for the research community.

Discussion

A shortcoming of the CIF is that, for the sake of being as comparable as possible, the generalized categories from the scheme do not consider survey-specific curation policies.[40]

A more fundamental weakness is that the index value, of course, depends on the initial quality level curators are confronted with. If the data quality is excellent from the start, and no errors are found, the CIF is low, despite all efforts taken. Therefore, it could be argued that the value of data curation lies in the curation process, independent of how many errors are found and corrected. To estimate merely the curation effort, Columns 3 and 4 (or 6) of the Data Curation Scheme (Table 1) can be used as a baseline.[41] The CIF, however, goes beyond this by identifying and quantifying the curation effect. This paper shows that even long-standing survey programs, where generally high data quality can be assumed because national teams are well-trained and have access to thoroughly developed method guidelines, gain significantly in data quality through centralized data curation.

Apart from the pleasant side effect that the calculated values show the curators that their work makes a difference, the CIF provides an excellent argument to demonstrate the value of curation for the research community. In grant applications, for example, researchers tend to be reluctant to include budget requests for curation tasks because they doubt reviewers understand the workload and their benefit and, therefore, refrain from accepting the related financial resources required for good curation work. The CIF values provide a measure to justify this work and, thus, might increase the chances of being positively considered in grant applications for any survey.

Curators or researchers who are planning to calculate the CIF for their survey data can follow our example for designing a Data Curation Scheme or organize the error

[40] For an example see Footnote 19.

[41] A suitable approach would be, for example, a transformation of curation effort into the time spent and the quantification of this time into are score (Perry & Netscher, 2022).

categorization based on the survey's cleaning reports. If different surveys are going to be compared, as in our example, an ISSP and an EVS survey, to enable the process to be streamlined, the scheme should be designed before the curation work starts.

Another aspect that could be worth considering for calculating CIFs is their customization through weighting. In our example, the CIF is unweighted, which means it treats all curation acts the same, irrespective of how effective the measure is (e.g., recoding versus just documentation) and how consequential the error would be for analyses. A reversed scale, as shown in our newspaper headline example previously, probably has the potential of being more consequential than a wrongly coded missing value. Therefore, a customizing approach could be to classify error types according to their significance for the specific data and the analyses in specific research fields and to weight them accordingly.

Conclusion

Society has a right to valid scientific research results that can only be generated based on high-quality data. Lacking or superficial data curation, or data curation performed by laypersons, seriously affects data quality negatively. Beyond that, imprecise data might create follow-up mistakes for future waves of the same survey. Even if research communities would, of course, like to have access to collected data as quickly as possible, remember that in terms of data processing, "fast isn't always good." Thorough data curation takes time and resources, which are sometimes difficult to justify to stakeholders, who know data only from a user's perspective.

This paper presents an approach to quantify single curation steps and puts a value on the curation impact. To calculate a CIF, curation measures must be documented in a pre-structured template that is called a Data Curation Scheme. This scheme can be designed as in Table 1, but should be adjusted to the structure of any pre-existing cleaning reports if possible. The sum of the curation measures applied must be set in relation to the total number of items on each curation level: for international survey data, we recommend building the scheme on:

- the Case level;
- the Variable level;
- the Dataset level;
- the Documentation level.

If the curation impact for a different kind of data is explored, these benchmarks might need to be adjusted to the survey's characteristics. The summed result, the CIF, is the percentage of data improvement that can be reported. The information from the Data Curation Scheme helps the structural assessment of data quality variation between countries within a survey or, under appropriate prerequisites, the comparison across surveys. It can be reported to justify time and staff resources spent on data curation during a survey data life cycle, e.g., when calculating project budgets in grant applications.

Data Availability

The article is based on the European Values Study (EVS) Integrated Dataset (v4.0.0, 2020-10-20) and the International Social Survey Programme: ISSP 2013 - National Identity III dataset (Version 2.0.0), which include data from 34 countries and are available from GESIS (EVS, 2020; ISSP Research Group, 2015). In its final release (v5.0.0, 2022-05-16), the EVS Integrated Dataset includes data from 36 countries.

Acknowledgements

We would like to thank our colleagues at the GESIS department Survey Data Curation, especially Markus Quandt and Boris Heizmann, as well as two anonymous reviewers, for their valuable comments and suggestions on earlier versions of this article.

References

American Association for Public Research (AAPOR). (2023). Standard definitions. *Final dispositions of case codes and outcome rates for surveys.* (10th ed.). AAPOR. Retrieved from https://aapor.org/standards-and-ethics/standard-definitions/

Bechert, I., Christmann, P., Franken, A., Franzese, F., Koomen, M., Krell, K., Kunz, T., Pflüger, M., Scheller, P., Trommer, M., and von Rotz, C. (2023). Plausibilisierung: eine Handreichung zum Umgang mit inkonsistenten Daten in den Sozialwissenschaften. *Bausteine Forschungsdatenmanagement* (1): 1-84. doi: 10.17192/bfdm.2023.1.8413

Biemer, P. P., Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C., & West, B. T. (Eds.). (2017). *Total Survey Error in Practice* (1st ed.). Wiley. doi:10.1002/9781119041702

Brislinger, E., & Moschner, M. (2019). Datenaufbereitung und Dokumentation. In *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten* (pp. 97–114). Verlag Barbara Budrich. doi:10.3224/84742233

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating Longitudinal Mean and Covariance Structures Analysis (LMACS) and Multiple Indicator Latent Growth Modeling (MLGM). *Organizational Research Methods*, *1*(4), 421–483. doi:10.1177/109442819814004

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European Social Survey. *Survey Research Methods*, *2*, 33–46. doi:10.18148/SRM/2008.V2I1.365

de Jonge, T., Veenhoven, R., & Arends, L. (2014). Homogenizing responses to different survey questions on the same topic: Proposal of a scale homogenization method using a reference distribution. *Social Indicators Research*, *117*(1), 275–300. doi:10.1007/s11205-013-0335-6

European Values Study (EVS). (2020). *European Values Study 2017: Integrated Dataset (EVS 2017)* (ZA7500; 4.0.0) [Data set]. GESIS Data Archive. doi:10.4232/1.13560

German Council for Scientific Information Infrastructures (RfII). (2020). The Data Quality Challenge. Recommendations for Sustainable Research in the Digital Turn. https://rfii.de/?p=4203

GESIS. (2020). European Values Study (EVS) 2017: Methodological guidelines. *GESIS Papers*. doi:10.21241/SSOAR.70110

Grigat, S. (2017). Von Österreich lernen Die FPÖ als Vorbild der AfD und Antisemitismuskritik in Zeiten islamistischer Mobilmachung. In *AfD & FPÖ* (pp. 9–26). Nomos Verlagsgesellschaft mbH & Co. KG.

Groves, R. M. (1989). *Survey Errors and Survey Costs* (1st ed.). Wiley. doi:10.1002/0471725277

Groves, R. M. (Ed.). (2009). *Survey methodology* (2nd ed.). Wiley.

Heinisch, R., & Werner, A. (2019). Who do populist radical right parties stand for? Representative claims, claim acceptance and descriptive representation in the Austrian FPÖ and German AfD. *Representation*, *55*(4), 475–492. doi:10.1080/00344893.2019.1635196

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*, *3*(1), 134–140. doi:10.2218/ijdc.v3i1.48

Inter-university Consortium for Political and Social Research (ICPSR). (n.d.). *Data Management & Curation.* Retrieved 22 November 2023, from https://www.icpsr.umich.edu/web/pages/datamanagement/index.html

International Labour Office (ILO). (2012). International Standard Classification of Occupations 2008 (ISCO-08): Structure, Group Definitions and Correspondence Tables. International Labour Office.

ISSP Demographic Methods Group. (2024). *ISSP Background Variables Guidelines for ISSP 2022*. GESIS. https://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/issp/members/codinginfo/BV_guidelines_for_issp2022.pdf

International Social Survey Programme (ISSP) Research Group. (2015). *International Social Survey Programme: National Identity III - ISSP 2013 International Social Survey Programme: National Identity III - ISSP 2013* (Version 2.0.0) [dataset]. GESIS Data Archive. doi:10.4232/1.12312

Johnston, L. (2014). A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot. University Digital of Minnesota Conservancy. Retrieved from https://hdl.handle.net/11299/162338

Knight, M. (2017, December 25). What Is Data Curation? Retrieved from https://www.dataversity.net/what-is-data-curation/

Kuriakose, N., & Robbins, M. (2016). Don't get duped: Fraud through duplication in public opinion surveys. Statistical Journal of the IAOS, 32(3), 283–291. doi:10.3233/SJI-160978

Otto, F. (2019). Julia Partheymüller: 'Die Frage, wie rechts die FPÖ ist, stellt sich kaum noch jemand'. *Die Zeit*. Retrieved from https://www.zeit.de/politik/ausland/2019-09/wahl-oesterreich-rechtspopulismus-fpoe-afd-waehler

Perry, A. & Netscher, S. (2022), Measuring the time spent on data curation. *Journal of Documentation*, *78*(7), 282–304. doi:10.1108/JD-08-2021-0167

Rhoades, S. (1993). The Herfindahl-Hirschman index. *Federal Reserve Bulletin*, *79*, 188–189.

Smith, T. W. (2011). Refining the total survey error perspective. *International Journal of Public Opinion Research*, *23*(4), 464–484. doi:10.1093/ijpor/edq052

Smith, T. W. (2020). 6. Using the total survey error paradigm to improve cross-national research on religion. In R. Finke & C. D. Bader (Eds.), *Faithful Measures* (pp. 167–198). New York University Press. doi:10.18574/nyu/9781479875214.003.0009

UNESCO Institute for Statistics. (2012). *International standard classification of education: ISCED 2011*. Retrieved from https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33. doi:10.1080/07421222.1996.11518099

Weisberg, H. F. (2005). The total survey error approach: A guide to the new science of survey research. University of Chicago Press.

Appendix

R^2 = 0.205	b coefficients	Standard Error
*Parties**		
SPÖ	–0.34	0.07
Die Grünen	–0.82	0.11
NEOS	–0.52	0.16
FPÖ	0.50	0.08
Controls		
Gender	not significant	
Age	0.01	0.00
Education	–0.05	0.02
Born in AT	–0.22	0.10

*Reference (ÖVP Österreichische Volkpartei (Conservatives)). All coefficients but gender were significant at a 5% level. SPÖ = Sozialdemokratische Partei Österreichs (Social Democrates), Die Grünen (Greens), NEOS = Das Neue Österreich und Liberales Forum (Liberals), FPÖ = Freiheitliche Partei Österreichs (Nationalists); AT=Austria

Figure A1: Results from an OLR regression. Dependent variable Austria, EVS 2017, v184, before correction. The higher the b coefficient, the more positive attitudes toward immigrants.