HANDBOOK OF COMPUTATIONAL SOCIAL SCIENCE, VOLUME 1

Theory, Case Studies and Ethics

Edited by Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu, and Lars Lyberg

First published 2022

ISBN: 978-0-367-45653-5 (hbk)
ISBN: 978-0-367-45652-8 (pbk)
ISBN: 978-1-003-02458-3 (ebk)

9

CAUSAL AND PREDICTIVE MODELING IN COMPUTATIONAL SOCIAL SCIENCE

Uwe Engel

DOI: 10.4324/9781003024583-10

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

9

CAUSAL AND PREDICTIVE MODELING IN COMPUTATIONAL SOCIAL SCIENCE

Uwe Engel

Introduction

Computational social science (CSS) is best understood as an interdisciplinary field in the intersection of data science and social science that pursues causal and predictive inference as its two main objectives. With roots in mathematical modeling and social simulation in sociology, in particular the increasing digitization of life turned CSS into a dynamically developing scientific field. The more human life takes place in digital environments, the more behavioral data of interest to the social and behavioral sciences accumulate. These data are only available digitally and thus require proper "new computational" methods of data collection, data management, data processing, and data analysis (Lazer et al., 2020). Delimitations of CSS reference big data, computational methods, and data science facets, as detailed in the following.

Perspectives on computational social science methodology

Computational methods in sociology

The impression of an inflationary use of the adjective "computational" is certainly justifiable. As Hox (2017, p. 3) states, with explicit consideration of the social and behavioral sciences, it seems that almost any scientific field can now be preceded by the adjective "computational". Though it is true that computational methods have reached an extremely high significance, it is equally true that the social sciences have developed computational methods from the outset: statistical data analysis and mathematical modeling are by no means new computational tools in this field. Coleman's (1964) seminal *Mathematical Sociology* is one of the best exemplifications of the precursors of contemporary analytical sociology. Other examples are the Columbia School of Social Research and its pioneering development work on multivariate analysis (e.g., Lazarsfeld, 1955; Rosenberg, 1968), the mathematical study of change (Coleman, 1968), methods of replication (Galtung, 1969), prediction studies (e.g., Goodman, 1955), the analysis of change through time by panel analysis (e.g., Lazarsfeld, Berelson, & Gaudet, 1955), Lazarsfeld's latent structure analysis (McCutcheon, 1987), longitudinal data analysis (Coleman, 1981), and

DOI: 10.4324/9781003024583-10

multilevel methodology (e.g., Lazarsfeld & Menzel, 1969 [1961]; Coleman et al., 1966; Coleman, 1990). The "social simulation modeling" area in CSS and its multilevel approach towards social complexity as detailed in Cioffi-Revilla (2017, pp. 12–20, 205ff., 331ff., 470ff.) is also highlighted.

Hox (2017, p. 3) refers to the simulation branch when reviewing "three important elements" of CSS, namely "big data, analytics, and simulation". In his view, CSS refers "to the application of computational methods to explore and test scientific (social, psychological, economic) theories", the field being "often interpreted as equivalent to the use of big data in social science. However, although computational social science relies strongly on big data and the ability to analyze these, it also includes computer simulation and text analysis". Edelmann, Wolff, Montagne, and Bail (2020, p. 62), too, highlight the simulation branch in CSS. They point out that "the term computational social science emerged in the final quarter of the twentieth century within social science disciplines as well as science, technology, engineering, and mathematics (STEM) disciplines". They continue by writing that within social science, "the term originally described agent-based modeling – or the use of computer programs to simulate human behavior within artificial populations", whereas within STEM fields, by contrast, "any study that employs large datasets that describe human behavior" is "often described as computational social science".

Big data

Shah, Capella, and Neuman (2015, p. 7) views CSS

> as a *specific subcategory* of work on big data. It is an approach to social inquiry defined by (1) the use of large, complex datasets, often – though not always – measured in terabytes or petabytes; (2) the frequent involvement of "naturally occurring" social and digital media sources and other electronic databases; (3) the use of computational or algorithmic solutions to generate patterns and inferences from these data; and (4) the applicability to social theory in a variety of domains from the study of mass opinion to public health, from examinations of political events to social movements.
>
> (emphasis added)

Alvarez (2016, p. 5) stresses in his introduction to *Computational Social Science* that it is

> not about "big data". . . . Instead, the focus is on the methodological innovations driven by the availability of *new types of data* or by *data of a different scale* than has been previously possible. . . . [T]he emphasis rather is on the cross-disciplinary applications of statistical, computational, and machine learning tools to the new types of data, at a larger scale, to learn more about politics and policy.
>
> (emphasis added)

Data science

In line with reference to terms such as "analytics" and "statistical, computational, and machine learning tools", one can delimit CSS also by reference to data science. Hox (2017, p. 3), for instance, regards CSS as an "interdisciplinary field that includes mathematics, statistics, data science, and, of course, social science". Proponents of data science even tend to subsume CSS under this discipline. "Data science" is then regarded as a "new interdisciplinary field that

synthesizes and builds on statistics, informatics, computing, communication, management, and sociology" . . . "to transform data to insights and decisions" (Cao, 2017, pp. 43, 8). Provost and Fawcett (2013, p. 4f.) "view the ultimate goal of data science as improving decision-making, as this generally is of direct interest to business". Similarly, Kelleher and Tierney (2018, p. 1) stress as a commonality the focus on "improving decision making through the analysis of data" when they point out that the "terms *data science*, *machine learning*, and *data mining* were often used interchangeably". Data science is broader in scope, while machine learning "focuses on the design and evaluation of algorithms for extracting patterns from data" and "data mining generally deals with the analysis of structured data". Data science

> also takes up other challenges, such as the capturing, cleaning, and transforming of unstructured social media and web data; the use of big-data technologies to store and process big, unstructured data sets; and questions related to data ethics and regulation.
> (Kelleher & Tierney, 2018, p. 1f.)

Machine learning and artificial intelligence

The Royal Society (2017, p. 25) localizes machine learning at the intersection of artificial intelligence, data science, and statistics, with applications in robotics. In computer science, deep learning is viewed as part of machine learning, and machine learning is viewed as part of artificial intelligence. Kelleher (2019, p. 4) dates the birth of AI to "a workshop at Dartmouth College in the summer of 1956". The Royal Society (2017, p. 28) outlines a similar picture in a timeline on developments in machine learning and AI by mentioning the Turing Test in 1950 and the Dartmouth Workshop. Part of this timeline also provides a clue that several key concepts in machine learning are derived from probability theory and statistics, the roots of which date to the 18th century (e.g., Bayes' theorem). James, Witten, Hastie, and Tibshirani (2013) use the term statistical learning and point to the essential role of statistical methods in the machine learning field. Ghani and Schierholz (2017, p. 148) write that "over the past 20 years, machine learning has become an interdisciplinary field spanning computer science, artificial intelligence, databases, and statistics". It is worth mentioning that the special relevance of machine learning derives from both the powerful capabilities of deep learning and artificial intelligence and their resulting impact on contemporary society and from the considerably less spectacular machine-learning applications in statistical data analysis.

Prediction and causal inference

In a similar vein, Kuhn and Johnson (2013, p. 1) refer to machine learning, artificial intelligence, pattern recognition, data mining, predictive analytics, and knowledge discovery and point out that "while each field approaches the problem using different perspectives and toolsets, the ultimate objective is the same: *to make an accurate prediction"*. Consequently, the authors pool "these terms into the commonly used phrase *predictive modeling*". Yarkoni and Westfall (2017, p. 2) draw attention to the "underappreciated tension between prediction and explanation", characterize current practice in psychology as "explanation without prediction", and raise substantial doubts about the predictive power and replicability of such explanatory research. "The crucial point is that prediction and replicability are the same problem", as Hindman (2015, p. 60f.) outlines; machine-learning-based approaches produce models that are more likely to be reproduced by other scholarship because they are more robust than standard practice and because "reducing the out-of-sample error produces more stable findings across different researchers

and different studies". Lin (2015) vigorously defends big data analysis against the reservation that this approach neglects theory. He argues, somewhat provocatively, that much of the controversy stems from a fundamental confusion about the purpose of big data. Using the example of commercial matchmaking services, he claims "that understanding is not necessary for prediction. . . . Whether any of their matchmaking algorithms are based on their current (academic) understanding of attraction is largely irrelevant" (Lin, 2015, p. 39).

This view draws criticism from proponents of causal analysis in social research. Latent variable modeling acts on the assumption that theoretical input is needed for both the measurement and structural part of a causal model. Yet not all explanations predict satisfactorily, and not all good predictions explain, as Troitzsch (2009) remarks. Hofman, Sharma, and Watts (2017, p. 486) "argue that the increasingly computational nature of social science is beginning to reverse (the) traditional bias against prediction", while highlighting the view that "predictive accuracy and interpretability must be recognized as complements, not substitutes, when evaluating explanations". Similarly, James et al. (2013, Chapter 2) elaborate on doing prediction and causal inference within a statistical learning approach. Balancing these two objectives in an analysis might be advantageous because prediction accuracy and generalizability cannot be maximized simultaneously. While maximizing R^2 is certainly suitable for increasing prediction accuracy, such a strategy runs the risk of overfitting and impairing the generalizability of found results. Data mining thus trusts techniques based on "replication" and "cross-validation", in which the use of training, tuning, and test samples guards against overfitting (Attewell & Monaghan, 2015, p. 14). The power of predicting unseen (out-of-sample) data can be regarded as a more relevant criterion than the size of a "theoretically privileged" regression coefficient or a model fit statistic (Yarkoni & Westfall, 2017, p. 2).

CSS, data science, and social science

It is constructive to locate CSS in the intersecting set of data science with social science and underlay the fields with possible data types, as shown in Figure 9.1. Data science is conceived in the broader sense outlined previously. Its core is the machine learning field, along with computational methods for collecting, cleaning, transforming, and processing datasets on even the largest scales (Luraschi, Kuo, & Ruiz, 2020).

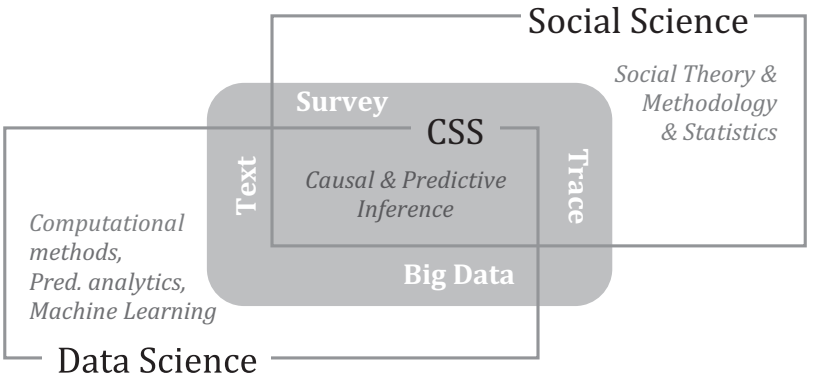

Figure 9.1 Causal and predictive inference in CSS

A substantial overlap exists between statistical methods employed in the machine-learning tradition, on the one hand, and the causal and descriptive analysis of observational studies in the counterfactual research tradition of social research, on the other hand (Morgan & Winship, 2007). Examples include principal component analysis (PCA), classification and clustering methods, and linear and logistic regression, as well as replication methods (e.g., James et al., 2013; Kuhn & Johnson, 2013; Lantz, 2019). Despite differences in technical language, both traditions share a subset of common statistical learning methods. In addition, both research traditions share the objective of inferring valid insights from nonexperimental data. This, in turn, requires tools for meeting two overwhelmingly important demands, namely the adequate ascertainment of validity and the attainment of effective control over possible sources of error.

In Figure 9.1, "Pred. analytics" stands for predictive analytics, to indicate the great importance attached to prediction and the achievement of prediction accuracy in the machine-learning framework. However, following James et al. (2013, Chapter 1), the basic possibility of achieving trade-offs between prediction accuracy and model interpretability is also regarded as a key feature of the statistical learning approach to data science. It is exactly this feature that opens the opportunity of doing both predictive and causal inference within a joint data-science framework and thus the opportunity to counterbalance the susceptibility to prediction bias in this field.

On the one hand, particular strengths of the statistical learning field include highly flexible modeling options, capabilities of handling nonlinear relationships, and tools for assessing validity on a regular basis. On the other hand, in the counterfactual and statistical research tradition pursued in the social and behavioral sciences, much effort is invested in latent variable modeling and the development of methodological and statistical tools for coping effectively with potential sources of error, such as confounding factors, selective unit and item nonresponse, and measurement effects. Along with the existing overlap of statistical methods, a data science that capitalizes on the best of these two worlds appears thus particularly constructive. The same is true for a CSS that takes advantage of import from such a data science, "even – or rather especially – when their datasets are small" (Hindman, 2015, p. 49), and the social and behavioral theories and methodologies that come from the social and behavioral sciences.

Big data, digital textual and behavioral trace data, and survey data

Volume (scale of data), velocity (analysis of streaming data), variety (different forms of data), and veracity (uncertainty of data) represent the original four Vs of big data (IBM, 2012). To this, add the fifth and sixth Vs, virtue (ethics) and value (knowledge gain) (Quan-Haase & Sloan, 2017, pp. 5–6). All these Vs make great demands on data collection and analysis. Volume matters: Though size is no inherent data quality criterion, large datasets facilitate multivariate analysis and validation but can make analysis more complicated (statistical inference, distributed computing). The collection and analysis of streaming data may also require special computational tools in the shape of online learning algorithms (see Ippel et al., Chapter 5 in Volume 2 of this handbook). In CSS, variety is a challenge through the simultaneous use of quite different kinds of data, such as the user-generated content in social media and the digital marks people leave when surfing the web. Such textual and digital-trace data complement survey data and make great demands on their separate and combined analysis. Veracity refers to the whole spectrum of data-quality criteria as, for instance, known in social-science methodology. A whole range of issues and challenges is accordingly involved in this regard. These issues become evident when

transferring the total error paradigm known from survey methodology to social media research (Biemer, 2017; Quan-Haase & Sloan, 2017; Franke et al., 2016; see also Sen et al., Chapter 9 in Volume 2 of this handbook).

Against the background that a major strand of CSS research develops in the highly frequented social-media context, the perceived close link between CSS and big data reported previously comes as no surprise. However, even if both behavior and behavioral marks in digital environments cover the major focus of social research in CSS, survey data play an important role too. Big data analysis benefits from both the availability of huge pools of compoundable survey data (e.g., Warshaw, 2016, p. 28ff.) and the data quality of surveys.

Causal inference in social research

Morgan and Winship (2007) describe the counterfactual model for observational data analysis, also known as the "potential outcomes framework" (Salganik, 2018, pp. 62–70), in greater detail. In doing so, they refer extensively to Pearl's (2000) work on graphical representations of causal relationships (graph theory). The authors deal particularly with three basic strategies for estimating causal effects: (1) to condition on variables "that block all backdoor paths from the causal variable to the outcome variable", (2) the use of "exogenous variation in an appropriate instrumental variable to isolate covariation in the causal and outcomes variables", and (3) the establishment of an "isolated and exhaustive mechanism that relates the causal variable to the outcome variable and then calculates the causal effect as it propagates through the mechanism" (Morgan & Winship, 2007, p. 26). The authors present a selective history of the use of experimental language in observational social science and regard the counterfactual model of causality as valuable because it helps researchers "to stipulate assumptions, evaluate alternative data analysis techniques, and think carefully about the process of causal exposure" (Morgan & Winship, 2007, p. 7). They also refer to "equation-based motivations of observational data analysis" to draw attention to the weaknesses of a "naïve usage of regression modeling", path analysis, and structural equation modeling as tools for the estimation of causal effects. They conclude that the rise of regression has "led to a focus on equations for outcomes, rather than careful thinking about how the data in hand differ from what would have been generated by the ideal experiments one might wish to have conducted" (Morgan & Winship, 2007, p. 13).

Some caution is advisable, indeed, if regression analysis is used for causal analysis. Bollen (1989, Chapter 3) uses three criteria to infer cause: isolation, association, and direction of causation (establishing causal priority). In doing so, the crucial point is that "isolation" designates actually a "pseudo-isolation" condition realized by means of assumptions about the error term of a regression equation (i.e., that the error is not correlated with the explanatory variable). Such assumptions are most important to get the model closed and thereby get changes in y attributable to corresponding changes in the explanatory variable(s). The clue that regression diagnostics are available to test such assumptions on a routine basis is relevant. Furthermore, "theory" plays an important role in both the counterfactual model ("the need to clearly articulate assumptions grounded in theory, that is believable", Morgan & Winship, 2007, p. 13) and structural equation modeling. The need for basing causal models in substantial theory is challenging in view of the basic limitations of model testing procedures. In the end, goodness-of-fit can only prove the compatibility of data and model implications, while it cannot rule out alternative models that have the same implications as a given tested model (Galtung, 1969; Stinchcombe, 1968, p. 19f.). Furthermore, as Pearl (2012, p. 68) notes, there exists a "huge logical gap between "establishing causation", which requires careful manipulative experiments, and "interpreting parameters as

causal effects", which may be based on firm scientific knowledge or on previously conducted experiments. . ., to conclude that "as a matter of fact. . ., causal effects in observational studies can only be substantiated from a combination of data and untested theoretical assumptions, not from the data alone".

Notwithstanding such basic limitations, "causal models" have prominent roots in the scientific work on path analysis and simultaneous-equation techniques realized in the 1950s and 60s, as Blalock's (1971) collection of such work documents. Since then, structural equation modeling and, more generally, latent variable (mixture) modeling evolved into a highly advanced tool for the statistical analysis of observational data; with strengths in controlling for (1) random and systematic measurement error, (2) methods effects by MTMM modeling (e.g., Marsh & Grayson, 1995), and (3) data missing at random and data not missing at random, by multiple imputation and selection/pattern mixture modeling (e.g., Winship & Morgan, 1999; Enders, 2010, van Buuren, 2018).

From the outset, longitudinal modeling using panel data played an essential role (e.g., Blalock, 1985; Coleman, 1981; Hagenaars, 1990), among other things, in securing by this repeated measurement design the validity of a key assumption of causal inference: that the alleged cause must precede the effect. As de Vaus (2001, p. 34 ff.) outlines, it must make sense to infer cause from covariation; he notes three criteria for the necessary plausibility in this respect: (1) time order (cause must come before the effect; see Bollen (1989, p. 61f.) for a discussion of this criterion), (2) the dependent variable must be capable of change, and (3) the causal assertion must make sense theoretically. A panel design is basically constructive in this regard because causal order may be derived from the temporal order between the panel waves. Longitudinal modeling using this design changed somewhat from the classical cross-lagged panel model to latent growth curve modeling (e.g., Byrne, 2012). This latter design is not only particularly suitable for multi-wave panel data; this type of model avoids an inherent weakness of the classical, cross-lagged panel model (potential disparity in corresponding temporal and causal lag), performs causal analysis using concomitant (instead of residual) change as an indication of causality and realizes the combination of longitudinal and multilevel modeling (Engel & Meyer, 1996; Engel, Gattig, & Simonson, 2007).

Descriptive inference in large datasets

Specific scientific ends that large datasets tend to enable are the study of rare events, the study of heterogeneity, and the detection of small differences (Salganik, 2018, pp. 18–20).

McFarland and MacFarland (2015, p. 1) draw attention to statistical significance testing in case of big observational "found data" from social media and explain that the results in such cases "give analysts a false sense of security as they proceed to focus on employing traditional statistical methods" because hidden biases (for instance, population bias) would be easily "overlooked due to the enhanced significance of the results". As Salganik (2018, p. 20) put it in his elaboration on "ten common characteristics of big data": "While bigness does reduce the need to worry about random error, it actually *increases* the need to worry about systematic errors, the kind of errors . . . that arise from biases in how data are created". Hox (2017, p. 10) refers to cross-validation and raises the question of how much room is there at the bottom: "how well do these techniques work with 'small Big Data'." He points out that the formal testing of structural equation models becomes problematic in the case of usual sample sizes of comparative survey research because a small degree of misfit already rejects the model. The respective sensitivity of the likelihood ratio chi^2 test to sample size and the usual solution through an array of goodness-of-fit indices is well known (e.g., Byrne2012, pp. 66–77). Hox (2017, p. 10) continues by

stating that "a cross-validation or *k*-fold approach that analyses how well the structural equation model predicts actual scores in the holdout samples provides a much more direct test".

Validity of inference

Shadish, Cook, and Campbell (2002, p. 34) use "the term validity to refer to the approximate truth of an inference". They distinguish the four well-known *standard concepts of validity* and list threats pertaining to each of them (Shadish et al. (2002, pp. 37–39, 64ff.). Accordingly, statistical conclusion validity refers to inferences about the existence and size of covariation between treatment and outcome; internal validity reflects the approximate truth of an inference about whether an observed covariation between treatment and outcome reflects a causal relationship. Then, two further concepts cover aspects of the generalizability of inferences. While external validity designates the generalizability of inferences over variations in persons, settings, treatment, and outcomes, construct validity refers to inferences about what is being measured by outcome measurements and thus covers the generalizability to the higher-order constructs they represent. We can use these concepts as benchmarks for data-driven inference in CSS.

Threats to statistical conclusion validity

Shadish et al. (2002, pp. 45–52) present a longer list of threats to statistical conclusion validity. The reasons inferences about covariation between two variables may be incorrect include, among other things, two reasons with special relevance to big data analysis: violated assumptions of statistical tests and unreliability of measures. Clustered/nested data violate the assumption of independently distributed errors, while unreliability attenuates bivariate relationships. The suggested solution to the former threat is accordingly multilevel analysis; the suggested solution to the latter is latent variable modeling of observed measures.

Threats to internal validity

Threats to the internal validity of inferences may be due to the lack of longitudinal data that would allow for repeated measurements at the individual level. A possible consequence is an ambiguous temporal precedence of putative cause and effect, a well-known topic in cross-sectional data analysis. Beyond that aspect, mis-specified models may come along with biased estimates of causal effects. Examples include improper functional forms of regression equations, such as assumed linearity in the case of nonlinear relationships. Omitted and unneeded explanatory variables may also threaten internal validity and thus cause a need to check formal requirements using "regression diagnostics" (Kohler & Kreuter, 2017, pp. 290–310). Systematic measurement and nonresponse error are likely to cause biased estimates, too.

Threats to external validity

Low selection quality is certainly a threat in this regard. If units of analysis are selected from a frame, selection quality depends on frame quality, sampling mode, and selective nonresponse. This implies that even $N = ALL$ analyses may be impaired, as in the case of the "near-census projects" referred to by Lazer and Radford (2017, p. 29). In survey methodology, the quality of sampling frames matters, for instance, in the case of (online) access panels (Engel, 2015). In digital media research, the corresponding question is who uses a platform for what reasons and who does not? The generalizability of inferences over different platform-specific user populations, or

the generalizability to the general population, depends accordingly on platform-specific population composition. Ruths and Pfeffer (2014, p. 1064) exemplify this point by explicating, "the ways in which users view Twitter as a space for political discourse affects how representative political content will be" (similarly, Kim et al., 2014, p. 1984). As Lazer and Radford (2017, p. 29) put it: "big data often is a census of particular, conveniently accessible social world. All of Twitter is a census of Twitter. Data from Kickstarter are a census of Kickstarter". In addition, sampling mode (random vs. nonprobability) matters, as do possible selection effects due to technical (API) restrictions, privacy settings, guards against tracking, and the potential requirement of informed consent in digital forms of data collection (e.g., Keyling & Jünger, 2016, pp. 188–191). Then, selective nonresponse is likely to impair the generalizability – and thus the external validity – of sample-based inference to the very population from which the sample has been drawn.

Threats to construct validity

The "promise of big data collection is that expensive data collections are replaced by less costly 'found' data, and that sampling is replaced by analyzing all existing data: $N = All$, avoiding sampling error" (Hox, 2017, p. 8). On the one hand, this implies a waiving of carefully designed survey measurement instruments and, in turn, a waiving of any planned conceptualization and measurement of higher-order constructs. Mahrt and Scharkow (2013, p. 27) discuss "data-driven rather than theory-driven operationalization" and "availability bias" and conclude that researchers "should be aware that the most easily available measure may not be the most valid one, and they should discuss to what degree its validity converges with that of established instruments". Qiu, Chan, and Chan (2018, p. 61f.) address the topic of measurement validity in cases where "second-hand data collected by others such as social media services or mobile phone companies" were used. Among other things, they discuss three issues that can introduce measurement errors: noise due to non-individual accounts, machine-generated coding, and the need to use proxies for variables of interest. On the other hand, the non-reactive nature of such found data appears advantageous. Shah et al. (2015, p. 7) refer to this topic of "unobtrusive measures" (Webb, Campbell, Schwartz, & Sechrest, 1966). They use the phrase "naturally occurring" data and contrast these data with data from surveys and experiments and the typically associated biases through "experimenter effects" and "self-report/social desirability". The argument is that error-prone self-reports on behavior obtained, for instance, in survey research, are that way replaceable by unnoticed direct observations of the behavior itself.

Yet the crucial point is that "motivated misreporting" (Tourangeau, Kreuter, & Eckman, 2015) is not confined to survey research. Lazer and Radford (2017, p. 23) refer to "big data archives" and point out that such archives offer in principle "measures of actual behaviors, as compared with self-reports of behaviors". They continue by stating that "generally, self-reported behavior is noisy, with a variety of systemic biases. For example, people systematically lie about everything from whether they voted to what their weight and height are". Even if motivated misreporting can be ruled out with respect to the research process itself, found data are not error free only because they were obtained unobtrusively. Salganik (2018, p. 24) illustrates this point using the example of self-reports on Facebook. The story is simply that in the first instance, a self-report is a self-report irrespective of its (survey vs. social media) context embedment. However, that embedment may distinguish between surveys that guarantee strict anonymity and posts in the social media that are published information, likely with hidden reference/target groups as addressees in mind.

Beyond that, the meaning of the behavior observed in digital environments is only decipherable on the basis of hypotheses. Social science methodology usually precludes the possibility of hypothesis-free observation. A core question is, then: "What is the theoretical validity and significance of the data?" Following Mahrt and Scharkow (2013, p. 29), this example may illustrate the point:

> The number of times a message gets forwarded ("retweeted") on Twitter, for instance, may show a certain degree of interest by users, but without looking at the content and style of a tweet, "interest" could stand for popularity and support, revulsion and outrage, or simply the thoughtless routines of Twitter usage behavior.

Generalizability and replicability

Generalizability is a topic that covers much more than the previous population aspect. Galtung (1969, p. 316f.) provided a quite undisguised view of this topic as he took the perspective, "given a hypothesis *H*, show me the set of conditions *C* under which it is tenable", where time, space, and social background often be the most important variables. "An effort to generalize, then, is an effort to try out the hypothesis for other points or regions in the condition-space". Given the hypothesis is then confirmed for this new value of one or more condition variables, "we have a clear case of generalization"; otherwise, we would have to increase the complexity of the hypothesis by "working the condition variable into the hypothesis". If generalization is understood that way, then generalization implies replicability, and then replication attempts are theoretically directable towards any relevant condition variable, including variables related to conceptualization, measurement, data collection, and data analysis. Following Galtung (1969, p. 437 ff.), the target of systematic replication may then be pursued to increase the degree of confirmation of hypotheses/propositions by decreasing "the tenability of the argument that the findings are artefacts of the method".

In a recent classification of forms of replication in quantitative social science, this variant relates to "robustness" if the original data are used and "generalization" if new data are used for replication attempts. This classification distinguishes four possible replication goals (Freese & Peterson, 2017, p. 152): verifiability to check if the results of an original study are reproducible if the same analysis is performed on the same data, robustness to check how far target findings are merely the result of analytic decisions if a reanalysis on the original data uses alternative specifications, repeatability to determine whether key results of a study can be replicated by applying the original procedures to new data, and generalization to check if similar findings may be observed consistently across different methods or settings. The latter clearly involves replication attempts across datasets, for instance, as are involved in checks of the generalizability of findings obtained in a training set to the "unseen data" of a related holdout dataset. This practice certainly makes a good case for a key element of data-science practice, the use of independent, out-of-sample validation and cross-validation schemata, respectively.

Sources of variation

Conceptualization and measurement

The analysis of data from nonexperimental social research forces the researcher into unavoidable decisions in relation to the three major stages of conceptualization, modeling, and validation. Because any of these decisions can theoretically affect the final score, Figure 9.2 regards these

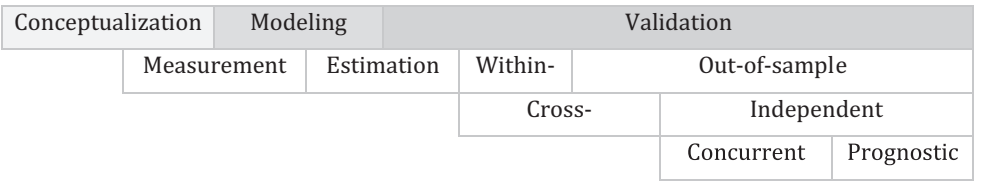

Figure 9.2 Conceptualization, modeling, and validation as sources of variation in nonexperimental research

deciding areas accordingly as possible sources of variation and presents them as an interleaved structure to mirror the relevant dependencies.

A first source of variation concerns the theoretical constructs of an analysis in relation to their supposed indicator variables. The general question is what indicator variable(s) best suit the theoretical concept(s). This question is particularly relevant in case the concept–indicator relation is not free of ambiguity. For instance, if voting behavior is used to indicate a person's inclination to vote for the extreme right, an ambiguous relation arises whenever relevant political parties consist of substantial currents of both moderate forces that adhere to the constitution of a country and extremist forces that try to overcome it. Then unambiguous coding of extremism is not possible using voting behavior as the sole indicator variable. This ambiguity is the more relevant the more popular is the party in the country in question. In today's sociological research, a related question is how far right a political party must be settled to pass the threshold from populism to extremism.

Measurement and modeling

Given a theoretical concept, the choice of suitable indicator variables is related to the question of how to bridge the levels of theoretical construct(s) and indicator variable(s) methodologically. Well-known options include the use of single- vs. multiple-indicator models and the explicit consideration of theoretical constructs by index formation, latent variable modeling, or a combination of both. Multiple-indicator models outmatch corresponding single-indicator models by a potentially better coverage of construct meaning. The inclination towards the political far right, for instance, might then be based on self-reports on voting behavior, the perceived closeness to political parties, and self-assessments on the common left-right scale. In addition, if constructs are related to eligible indicator variables by latent variable modeling and thus by testable correspondence hypotheses, then empirical testing replaces otherwise untested but believed assumptions. Similar advantageous features of the latent-variable framework are the checkability and control of random and systematic measurement error and the enabling of flexible modeling strategies within a wide array of options. Despite this principal flexibility, the focus remains on modeling perspectives that work with supposed functional forms instead of targeting the very functional form(s) that suit some relevant data best. James et al. (2013, p. 16 ff.) use a regression-like equation of the general form

$$Y = f\left(X\right) + \varepsilon$$

to introduce this distinction and a related localization of statistical learning methods in an analytical space, built to represent the trade-off between their flexibility vs. interpretability. A suitable choice of such methods then depends on whether an analysis seeks primarily descriptive/

causal inference, on the one hand, or unit-related prediction, on the other hand. Similarly, model design depends on these superordinate orientations. Simply to refer to the target complexity of models in terms of variables and functions: attempts at increasing the prediction accuracy of a model are likely to lead to comparably larger than smaller numbers of predictor/explanatory variables, while the reverse can be expected in the case of attempts at revealing, for instance, individual statistical/causal effects of particular interest. In addition, attempts at increasing prediction accuracy are likely to lead to models that estimate nonlinearities and statistical interactions often assumed/restricted to be zero in standard models for the estimation of linear and additive effects.

Modeling and estimation

The large variety of standard and robust estimation methods is well known to statisticians, data scientists, and structural equation modelers. Regularization is a standard topic in the machine-learning literature. The same holds true for the treatment of missing data caused by unit and item nonresponse in the statistical analysis of survey data. The proper treatment of clustered data structures by multilevel modeling is a further case in point. All such modeling/estimation options require decisions on how to proceed and, theoretically, any of these decisions can affect the final score. This fact calls for sensitivity analysis and systematic replication. Following recent suggestions (e.g., Harms, 2019; Ly, Etz, Marsman, & Wagenmakers, 2019), the computation of replication Bayes factors as an alternative to hypothesis testing in the frequentist tradition may be considered. Recently, the computation of Bayes factors was also proposed for the selection of replication targets (Field, Hoekstra, Bringmann, & van Ravenzwaaij, 2019).

Estimation and validation

Model estimation implies commonly used goodness-of-fit statistics (e.g., Byrne, 2012, pp. 66–77; Beaujean, 2014, pp. 153–166). The list includes the popular likelihood ratio chi^2 test statistic, incremental indices of fit (comparative fit index [CFI] and Tucker–Lewis index [TLI]), the root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMR), the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the mean square error (MSE), all of them for within-sample evaluations of model fit. Browne and Cudeck's (1993, p. 147 ff.) content validity index (CVI) and expected content validity index (ECVI) are two variants for cross-validation that also belong to the collection of common indices of fit. The CVI is the cross-validation index for two independent samples from the same population: a "calibration sample" and a "validation sample". The model is fitted to the observed calibration sample covariance matrix to determine to what extent the implied (estimated) covariance matrix fits the observed covariance matrix of the validation sample. Phrased in today's data-science parlance, the procedure relates a fitted covariance obtained in a training sample to the corresponding observed covariance matrix of a holdout sample. Then the ECVI designates the expected value of the cross-validation index. This value is estimated for both calibration and validation samples, "using the calibration sample alone" for computing it (Browne & Cudeck, 1993, p. 148). In this manner, the ECVI represents a substitute for a real cross-validation. The ECVI is subsumed under the parsimony indexes and regarded as a "single-sample approximation to the cross-validation coefficient obtained from a validation sample" (Beaujean, 2014, p. 160 f.). The index is used for model comparisons and is akin to the AIC in reflecting "the extent

to which parameter estimates from an original sample will cross-validate in future samples" (Byrne, 2012, p. 72).

Model fit is assessable in different ways. The researcher has a choice between different fit criteria, and the final score may vary depending on that decision. If the criteria suggest contrary decisions as to the acceptability of a model in question, the situation is comparable to the *p*-hacking strategy discussed in the literature and is as problematic as this strategy. Yarkoni and Westfal (2017, p. 4 f.), for instance, describe *p*-hacking as a practice of procedural overfitting, "of flexibly selecting analytical procedures based in part on the quality of the results they produce". Similarly, Freese and Peterson (2017, p. 155) trace the "abiding concern . . . that published findings represent a best-case scenario among all the arbitrary and debatable decisions made over the course of analyzing data" back to "p-hacking, in which a researcher runs different analyses until they find support for their preferred hypothesis". An ambiguous overall picture of fit criteria is a realistic scenario given the potential weaknesses of model testing, for instance, in the SEM context. The list includes impairments due to the use of chi^2-based statistics in large samples, fit criteria that enable only relative assessment across models, and acceptability thresholds on single indices that are only approximately settable in the end. Therefore, both single-fit criteria and the overall picture that emerges from reviewing the possible fit criteria for a given model in conjunction must be considered. Given ambiguous overall evaluations, it is certainly problematic to simply pitch selectively on the subset of criteria that confirm a model while neglecting other relevant criteria. Transferred to the regression context, a similarly ambiguous situation arises if an acceptable R^2 goes along with the diagnosis that assumptions about the model's error term are possibly violated. In this regard, it is clearly more constructive not to rely only on the MSE and the related R^2. The correct use of modeling frameworks (e.g., linear model, generalized linear model) and estimators (e.g., OLS, WLS, ML) presume sets of testable assumptions. Narrowing the spectrum of possible modeling alternatives to the ones that meet such requirements is a sensible choice.

Validation requires fixing relevant fit criteria and evaluation metrics, respectively, for in-sample and out-of-sample evaluation. Out-of-sample orientation is a core feature of data science and machine learning. Basic choices include cross-validation and independent (concurrent, prognostic/temporal) replication. Kelleher and Tierney (2018, pp. 145–148), Ghani and Schierholz (2017, pp. 173–180), and James et al. (2013, pp. 176–186) describe relevant foundations and validation schemata. In doing validation in an out-of-sample orientation, a key idea is to evaluate a model in question on "unseen data", that is, on data not used for building and maybe improving a model in question. The availability of different options that allow validation itself to become a possible source of variation of the final score is noteworthy.

A case study of political extremism

The present analysis uses pooled data from rounds 1 to 9 of the European Social Survey (ESS) and includes all countries with continuous participation over these rounds. For it, we pooled two original ESS data sources, the cumulative file for rounds 1 to 8 and the file for round 9 (ESS, European Social Survey Cumulative File, ESS 1–8 (2018) and European Social Survey Round 9 Data (2018)). This results in an overall sample size of 340,215 respondents. A detailed documentation of this analysis (design, R script, results, link to the data frame used for analysis) is published at https://github.com/viewsandinsights/inference. The target variable is affinity to the political far-right in Europe. Because the emphasis is placed on the far in far-right and thus on a relatively small population group in many EU countries, a large overall sample size is required for a survey analysis of these groups. The pooled ESS data offer this option along with

the additional advantage of enabling time-related studies of the political far-right in Europe. We refer to this analysis for addressing causal inference in a setting that uses (1) big (survey) data, (2) different validation strategies known from the machine-learning context (independent replication using training/test samples and cross-validation), and (3) systematic replication in the case of unavoidable decisions concerning data management and measurement (coding ambiguity, proxy variables, scope in concept specification, and missing data treatment).

Theoretical background

The analysis examines the affiliation to the political far-right as a function of prejudice formation, value orientation, and trust. It follows the general idea that prejudice formation is understandable as a response to competitive relations in open societies that prize freedom. Esses, Jackson, Dovidio, and Hodson (2005, p. 227f.) point out that concerns for group status produce competition between groups for material resources and for value dominance in which "competition exacerbates prejudice between groups". The theoretical expectation is that "in social contexts in which equality of opportunity is prized and social mobility is encouraged, perceived competition is likely" (Esses et al., 2005, p. 229), because then "members of even the lowest group are encouraged to put forth effort, to rise, and to demand their rights" (Allport, 1954/1979, p. 222). As Esses et al. (2005, p. 229) formulate this idea, "allowance, or encouragement, of upward social mobility makes evident the possibility of downward mobility, which, in the presence of a visually salient outgroup, generates competition and animosity". Accordingly, the expected "outcome of competition is stereotyped and/or overtly hostile attitudes toward competitor groups".

A second background element is inspired by previous research on right-wing authoritarianism. Only loosely following up the pioneering "Berkeley theory" of Adorno, Frenkel-Brunswik, Levinson, and Sanford's (1950) Authoritarian Personality, Altemeyer (1996) studied this phenomenon empirically in greater detail. He pictures authoritarians as characters who believe that proper authorities should be trusted and deserve obedience and respect. He refers to perceived established authorities as those "people in our society who are usually considered to have a general legal or moral authority over the behavior of others", usually one's parents, religious officials, civic officers (the police, judges, heads of governments), and superiors in military service (Altemeyer, 1996, pp. 8–12). However, the "perceived" established authorities may be others as well. As Altemeyer (1996, p. 9) explains, some "extremists may reject normal authorities who (it seems to them) have betrayed the real, fundamental established authority: for example (their perception of) God's will, or the Constitution. They often believe the government has been taken over by Jews, homosexuals, feminists, Communists, and so on. Such extremists are right-wing authoritarian in this context – "superpatriots" who see themselves as upholding traditional values but whose fear and self-righteousness hammer with such intensity that they rehearse for violence and may cross the line to violence itself". Altemeyer (1996, p. 10) stresses that he is "using "right-wing" in a *psychological* sense of submitting to the perceived authorities in one's life", meaning psychological "right-wingers in their support for those they were raised to believe were the legitimate authorities". This spotlights the psychological relationship individuals may have to the principle of authority and the adherence to this principle in the contexts of contemporary life.

Prejudice formation and the political far right

Figure 9.3 graphs the assumed possible effect structures schematically. The analysis is aimed at a decision as to the causal status of the prejudice-to-far-right path. We refer to the section "Causal

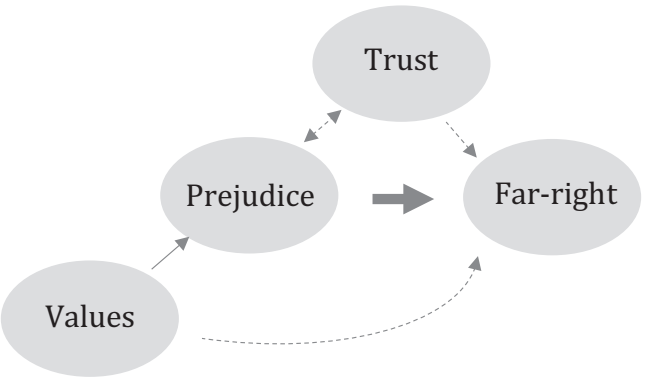

Figure 9.3 Does prejudice cause affinity to the political far-right?

Inference in Social Research" and check two criteria: (1) does *values* act as an instrumental variable whose effect on far-right affinity runs exclusively via prejudice formation, and (2) does *trust* act as an intervening (vs. antecedent) variable over which the effect of prejudice on far-right affinity is transmitted? These questions can be answered by checking the dashed arrows: it should be reasonable to assume that a potential direct effect of *values* on *far-right* drops to zero if *prejudice* is introduced or that *prejudice* affects *trust*, not vice versa, while *trust* affects *far-right* at the same time.

Validation in causal inference

As detailed elsewhere (https://github.com/viewsandinsights/inference), structural equation modeling is used to clarify the causal status of the prejudice-to-far-right path. Apart from the sociological interest in the subject matter, a series of systematic model comparisons was realized to explore model fit (chi^2, RMSEA, CFI, TLI, SRMR, ECVI) embedded in three basic validation strategies: (1) using a 70 to 30 percent random split of training and test sample, (2) using a systematic time-related split of training sample (data from ESS rounds 1 to 6) and test sample (data from rounds 7 to 9), and (3) using Browne and Cudeck's ECVI measure as a single-sample approximation to the cross-validation coefficient. Strategies (1) and (2) appear particularly important to take advantage of the causal graph approach referred to previously ("Causal Inference in Social Research"). A useful feature of this approach is its capability of deriving from a causal structure of effects and related checks for backdoor paths and instrumental and intervening variables the insight into which of the involved effects must be controlled for to avoid confounding and spurious estimates of effects. However, this moves causal inference close to circular reasoning: If a causal structure is already accepted as true in advance, there is no need to identify any such conditions. Then the model identifies causal effects properly, or it does not. However, if a model is only tentatively taken for granted and, if necessary, inductively elaborated on further in a training phase, then it makes much sense trying to validate such inductively improved models using the "unseen" data of a test sample.

Systematic replication in case of measurement alternatives

The political far right is approached in two ways: first, by the most-right category "10" on the 11-point, left-right scale ranging from zero to ten, thus by the respondents' own self-image

(sample estimate: 3.1% Europe-wide across all rounds and countries); second, by classifying pertinent political parties as far right following previous work in political science and additional sources. A detailed list of the country-specific coding sources used for this classification is part of the R script cited previously in "A Case Study of Political Extremism". Across all rounds and countries, 1.3% of the respondents expressed affinities ("feeling close" or "vote for") to these far-right parties. Combining both measures yields an estimate of 4.3% affinity either way. Both assessments are correlated with $r = 0.092$.

Worth mentioning in this connection is a practically unreducible ambiguity in the coding of right-wing parties as far right if they consist of currents of both loyal and hostile forces to a country's constitution. In the present analysis, this ambiguity created occasional scope for alternative coding decisions and, thus, subsequent variation of results.

Proxy variables

While the ESS data contain suitable indicator variables for the measurement of hostile attitudes towards immigrants, no direct indicator of the perceived competition in society is available. In lieu of this information, a proxy variable is used by acting on the auxiliary assumption that competitive relationships strengthen the motivational tendency to overreach others in the general pursuit of self-interest. The need for switching to proxy variables is certainly a limiting factor even when using an advanced causal analysis technique such as latent variable modeling at the same time. In the present case, this also applies to the measurement of right-wing authoritarianism by the human values scale. This scale is part of the standard ESS questionnaire program, which is an excellent instrument, although not specifically designed to measure the construct in question. It too is accordingly used as a proxy measure and exemplifies for the survey context a situation like the use of social media data in the CSS context: the use of proxy data because those data are available, even if more suitable measures are theoretically imaginable.

Concluding remarks

The social sciences are currently discovering machine learning (ML) as the new promise of replicable results because ML offers advanced statistical means and aims at balancing the predictive validity and generalizability of results. ML is clearly directed to predictive modeling and the target of maximizing prediction accuracy. Because too-accurately predicted findings run the risk of impaired generalizability to unseen data, at the same time, a guard against overfitting is realized by use of large datasets and validation through proving the replicability of results found in a training setting to the unseen data of holdout settings. Such a validation strategy comes along with a change of perspective from within-sample goodness-of-fit testing to assessing out-of-sample prediction error and bears basically on replication in relation to datasets. However, replication is a more general principle. Replication may also be targeted at other elements of a social research design/study, for instance, at the latent concepts, the observed variables, and all methods-related settings. This focus on replication has already allowed it to appear to be constructive to align CSS research with machine learning and the modeling, estimation, and evaluation tools therein. Statistical learning and the statistical analysis of observational data, as we practice it in social science research, have many statistical methods in common.

It would, however, be misleading and far from any sound balancing of explanation and prediction to abstain from causal inference in CSS. Causal modeling can look back on a long history in social research, and latent variable modeling implies powerful capabilities of dealing

with measurement error. The social and behavioral sciences aim centrally at understanding their scientific objects, and this understanding cannot go without appropriate theories about these objects. Hence, such theories are needed and must be subjected to proper empirical tests. However, irrespective of a possible focus on within- or out-of-sample validation, such tests can only help to decide if the implications of a theoretical model appear sufficiently consistent with some data in hand. This fact sets any such testing boundaries because even if such a model is proved consistent in this regard, no alternative theory/model with the same implications as the tested model is ruled out. This indefiniteness clearly weakens the theoretical significance of any such testing unless we conduct crucial tests in which a tested theory implies the negation of implications of some alternative theory.

References

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality*. Harper and Row.

Allport, G. W. (1979). *The nature of prejudice* (25th anniversary ed.). Basic Books (Original work published 1954).

Altemeyer, B. (1996). *The authoritarian specter*. Harvard University Press.

Alvarez, R. M. (2016). Introduction. In R. M. Alvarez (Ed.), *Computational social science. discovery and prediction* (pp. 1–24). Cambridge University Press.

Attewell, P., & Monaghan, D. B. (2015). *Data mining for the social sciences*. University of California Press.

Beaujean, A. A. (2014). *Latent variable modeling using R*. Routledge.

Biemer, P. P. (2017). Errors and inference. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science* (pp. 265–297). CRC Press.

Blalock, H. M. (Ed.). (1971). *Causal models in the social sciences* (1st ed.). Aldine Publ. Co.

Blalock, H. M. (Ed.). (1985). *Causal models in panel and experimental designs*. Aldine Publ. Co.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

Byrne, B. M. (2012). *Structural equation modeling with Mplus. Basic concepts, applications, and programming*. Routledge.

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys*, *50*(3), 43.1–43.42. https://doi.org/10.1145/3076253

Cioffi-Revilla, C. (2017). *Introduction to computational social science. Principles and applications* (2nd ed.). Springer.

Coleman, J. S. (1964). *Introduction to mathematical sociology*. The Free Press of Glencoe.

Coleman, J. S. (1968). The mathematical study of change. In H. M. Blalock, & A. B. Blalock (Eds.), *Methodology in social research* (pp. 428–478). McGraw-Hill.

Coleman, J. S. (1981). *Longitudinal data analysis*. Basic Books.

Coleman, J. S. (1990). *Foundations of social theory*. Belknap Press of Harvard University Press.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, M. D., & York, R. L. (1966). *Equality of educational opportunity*. US Government Printing Office.

De Vaus, D. (2001). *Research design in social research*. Sage.

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology, 46*, 61–68. https://doi.org/10.1146/annurev-soc-121919-054621

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Engel, U. (2015). Response behavior in an adaptive survey design for the setting-up stage of a probability-based access panel in Germany. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving survey methods: Lessons from recent research* (pp. 207–222). Routledge.

Engel, U., Gattig, A., & Simonson, J. (2007). Longitudinal multilevel modelling: A comparison of growth curve models and structural equation modelling using panel data from Germany. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Longitudinal models in the behavioral and related sciences* (pp. 295–314). Lawrence Erlbaum Associates.

Engel, U., & Meyer, W. (1996). Structural analysis in the study of social change. In U. Engel & J. Reinecke (Eds.), *Analysis of change. Advanced techniques in panel data analysis* (pp. 221–252). De Gruyter.

Esses, V. M., Jackson, L. M., Dovidio, J. F., & Hodson, G. (2005). Instrumental relations among groups: Group competition, conflict, and prejudice. In J. F. Dovidio, P. Glick, & L. A. Rudman (Eds.), *On the nature of prejudice. Fifty years after Allport* (pp. 227–243). Blackwell Publishing.

European Social Survey Cumulative File, ESS 1–8. (2018). *Data file edition 1.0. NSD – Norwegian Centre for research data, Norway – Data archive and distributor of ESS data for ESS ERIC*. doi:10.21338/NSD-ESS-CUMULATIVE

European Social Survey Round 9 Data (2018). *Data file edition 2.0. NSD – Norwegian Centre for research data, Norway – Data archive and distributor of ESS data for ESS ERIC*. doi:10.21338/NSD-ESS9-2018

Field, S. M., Hoekstra, R., Bringmann, L., & van Ravenzwaaij, D. (2019). When and why to replicate: As easy as 1, 2, 3? *Collabra: Psychology, 5*(1), 46, 1–15. https://doi.org/10.1525/Collabra.218

Franke, B., Plante, J.-F., Roscher, R., Lee, E.-S. A., Smyth, C., Hatefi, A., . . . Reid, N. (2016). Statistical inference, learning and models in big data. *International Statistical Review*, *84*(3), 371–389, https://doi.org/10.1111/insr.12176

Freese, J., & Peterson, D. (2017). Replication in social science. *Annual Review of Sociology, 43*, 147–165.

Galtung, J. (1969). *Theory and methods of social research* (1st ed., 1967). Universitetsforlaget.

Ghani, R., & Schierholz, M. (2017). Machine learning. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science* (pp. 147–186). CRC Press.

Goodman, L. A. (1955). Generalizing the problem of prediction. In P. F. Lazarsfeld & M. Rosenberg (Eds.), *The language of social research* (pp. 277–283). The Free Press.

Hagenaars, J. A. (1990). *Categorical longitudinal data. Log-linear, panel, trend, and cohort analysis*. Sage.

Harms, C. (2019). A Bayes factor for replications of ANOVA results. *The American Statistician*, *73*(4), 327–339, https://doi.org/10.1080/00031305.2018.1518787

Hindman, M. (2015). Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 48–62. https://doi.org/10.1177/0002716215570279

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488. https://doi.org/10.1126/science.aal3856

Hox, J. J. (2017). Computational social science methodology, anyone? *Methodology*, 13, 3–12. https://doi.org/10.1027/1614-2241/a000127

IBM Big Data & Analytics Hub. (2012). *The four V's of big data*. Retrieved December 30, 2020, from www.ibmbigdatahub.com/infographic/four-vs-big-data

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Kelleher, J. D. (2019). *Deep learning*. The MIT Press.

Kelleher, J. D., & Tierney, B. (2018). *Data science*. The MIT Press.

Keyling, T., & Jünger, J. (2016). Observing online content. In G. Vowe, & P. Henn (Eds.), *Political communication in the online world. Theoretical approaches and research designs* (pp. 183–200). Routledge.

Kim, A., Murphy, J., Richards, A., Hansen, H., Powell, R., & Haney, C. (2014). Can tweets replace polls? A U.S. health-care reform case study. In C. A. Hill, E. Dean, & J. Murphy (Eds.), *Social media, sociality, and survey research* (pp. 1638–2077). Wiley.

Kohler, U., & Kreuter, F. (2017). *Datenanalyse mit Stata [Data Analysis with Stata]* (5th ed.). De Gruyter Oldenbourg.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Lantz, B. (2019). *Machine learning with R. Expert techniques for predictive modeling*. Packt Publishing.

Lazarsfeld, P. F. (1955). Interpretation of statistical relations as a research operation. In P. F. Lazarsfeld & M. Rosenberg (Eds.), *The language of social research. A reader in the methodology of social research* (pp. 115–125). The Free Press.

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1955). The process of opinion and attitude formation. In P. F. Lazarsfeld & M. Rosenberg (Eds.), *The language of social research* (pp. 231–242). The Free Press.

Lazarsfeld, P. F., & Menzel, H. (1969). On the relation between individual and collective properties. In A. Etzioni (Ed.), *A sociological reader on complex organizations* (pp. 449–516). (1st ed., 1961). Holt, Rinehart and Winston.

Lazer, D. M. J., Pentland, A., Watts, D. J, Aral, S., Athey, S., Contractor, N., . . . Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062. https://doi.org/10.1126/science.aaz8170

Lazer, D. M. J., & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology, 43*, 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

Lin, J. (2015). On building better mousetraps and understanding the human condition: Reflections on big data in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 33–47, https://doi.org/10.1177/0002716215569174

Luraschi, J., Kuo, K., & Ruiz, E. (2020). *Mastering spark with R*. O'Reilly Media, Inc.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods, 51*, 2498–2508. https://doi.org/10.3758/s13428-018-1092-x

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20–33, https://doi.org/10.1080/08838151.2012.761700

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (p. 177–198). Sage.

McCutcheon, A. L. (1987). *Latent class analysis*. Sage.

McFarland, D. A., & McFarland, H. R. (2015). Big data and the danger of being precisely inaccurate. *Big Data & Society*, 1–4, https://doi.org/10.1177/2053951715602495

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference. Methods and principles for social research*. Cambridge University Press.

Pearl, J. (2009). *Causality. Models, reasoning, and inference* (1st ed., 2000). Cambridge University Press.

Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). Guilford Press.

Provost, F., & Fawcett, T. (2013). *Data science for business. What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Qiu, L., Chan, S. H. M., & Chan, D. (2018). Big data in social and psychological science: Theoretical and methodological issues. *Journal of Computational Social Science, 1*, 59–66. https://doi.org/10.1007/s42001-017-0013-6

Quan-Haase, A., & Sloan, L. (2017). Introduction to the handbook of social media research methods: Goals, challenges and innovations. In L. Sloan & A. Quan-Haase (Eds.), *The Sage handbook of social media research methods* (pp. 1–9). Sage.

Rosenberg, M. (1968). *The logic of survey analysis*. Basic Books.

The Royal Society. (2017). *Machine learning: The power and promise of computers that learn by example*. Retrieved December 30, 2020, from https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064. https://doi.org/10.1126/science.346.6213.1063

Salganik, M. J. (2018). *Bit by bit. Social research in the digital age*. Princeton University Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Co.

Shah, D. V., Capella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13, https://doi.org/10.1177/0002716215572084

Stinchcombe, A. L. (1968). *Constructing social theories*. The University of Chicago Press.

Tourangeau, R., Kreuter, F., & Eckman, S. (2015). Motivated misreporting: Shaping answers to reduce survey burden. In U. Engel (Ed.), *Survey measurements. Techniques, data quality and sources of error* (pp. 24–41). Campus.

Troitzsch, K. G. (2009). Not all explanations predict satisfactorily, and not all good predictions explain. *Journal of Artificial Societies and Social Simulation, 12*(1). Retrieved December 30, 2020, from http://jasss.soc.surrey.ac.uk/12/1/10.html

Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press.

Warshaw, C. (2016). The application of big data in surveys to the study of elections, public opinion, and representation. In R. M. Alvarez (Ed.), *Computational social science. Discovery and prediction* (pp. 27–50). University Press.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (2000). *Unobtrusive measures* (1st ed., 1966). Sage Classics 2. Sage.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*, 659–706. https://doi.org/10.1146/annurev.soc.25.1.659

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1–23. https://doi.org/10.1177/1745691617693393