Understanding data fabrication: Qualitative Comparative Analysis (QCA) of fabrication strategies

CHJ Hartgerink, Jan G Voelkel, Jelte M Wicherts, Marcel ALM van Assen

30 April, 2018

Introduction

Cases of data fabrication in research often peak the interest of people in research and beyond, where it speaks to the imagination to understand why, what, or how data were fabricated. For reasons why researchers fabricate data, some look at systemic origins (i.e., the bad barrel argument), such as the highly competitive research system (**???**, **???**), whereas others look to personality traits that might be predictive of likelihood to commit misconduct (i.e., the bad-apple argument; **???**). What was fabricated (i.e., which results) is often a question that drives scientific integrity committees established to investigate the case in order to correct the scientific record.

How data are fabricated remains puzzling because of incomplete knowledge of cases and lack of first-hand information on data fabrication. We have incomplete knowledge about data fabrication strategies because those that are effective at avoiding detection are not available for self-evident reasons (i.e., discovery bias). Vice versa, the cases we do know about can teach us what strategies are unsuccessful, but not the range of strategies that potentially are applied. However, those cases that are discovered often do not result in confessions with explicit descriptions of how they fabricated the data. For example, Diederik Stapel confessed to fabricating data and wrote a book about his recollections and interpretations (Stapel, 2012). This is an exceptional case because of the confessive approach he takes. However, even in this exceptional case little specific information is available as to how he fabricated data to be of value for research on data fabrication strategies. The following excerpt includes a general description of how he fabricated data over the years (Borsboom, 2013; Stapel, 2012):

> I preferred to do it at home, late in the evening, when everyone was asleep. I made myself some tea, put my computer on the table, took my notes from my bag, and used my fountain pen to write down a neat list of research projects and effects I had to produce … Subsequently I began to enter my own data, row for row, column for column … 3, 4, 6, 7, 8, 4, 5, 3, 5, 6, 7, 8, 5, 4, 3, 3, 2. When I was finished, I would do the first analyses. Often, these would not immediately produce the right results. Back to the matrix and alter data. 4, 6, 7, 5, 4, 7, 8, 2, 4, 4, 6, 5, 6, 7, 8, 5, 4. Just as long until all analyses worked out as planned. (p. 167)

Moreover, when scientific integrity committees investigate for data fabrication, rarely will they be able to conclusively state how data were fabricated. Some data fabrication strategies might be obvious if the raw data are available (e.g., copy-pasting responses), whereas others will be less obvious (e.g., multivariate modeling of the observed variables for the desired outcomes). Moreover, many fabricated results have underdetermined data fabrication strategies (i.e., multiplicity). That is, many different fabrication strategies can result in the observed data set, such that differentiating which one actually occurred based on the data is non-trivial (see also **???**). Additionally, given that odds of data availability decrease each year after publication (**???**), methods to learn about how data were fabricated retrospectively by looking at the dataset are increasingly unfeasible even if underdetermination was not a problem. Data availability is especially problematic in cases where it takes longer to uncover problems in the first place, such that data are more likely to be unavailable for investigation in cases that go undetected longer. Second-hand information about data fabrication offers relatively little indication on how data are actually fabricated by researchers due to this underdetermination and increasing lack of data availability.

Hence, first-hand knowledge from controlled settings about how researchers fabricate data is useful to further understand and maybe even improve detection of data fabrication in uncontrolled settings. There is unknown variability in how researchers fabricate data, which could result in foregone detection mechanisms if we focus on the limited and preselected knowledge that is available. However, we currently do not even know what methods are used and therefore only operate from hunches and inferences from psychology theory (e.g., Haldane, 1948) and how often they occur. As far as we know, only one such study asked participants to fabricate data, but did not investigate how they did so (**???**). Qualitative information would provide a first-hand insight into how data are fabricated and could provide fruitful avenues for the development of new statistical tools to detect data fabrication.

A better understanding of how researchers fabricate data can assist in developing- and validating tools to detect data fabrication. Previously, researchers developed methods to detect data fabrication in situ. For example, terminal digit analysis was developed as part of the Imanishi-Kari case [@]; testing for excessive amounts of high *p*-values was developed as part of the Fuji case [@]; variance analysis was developed as part of the Smeesters and Sanna cases (**???**). As such, it seems plausible that development of additional statistical tools can be facilitated by detailed descriptions of data fabrication cases. Validation of such statistical tools is also benefited by detailed descriptions of data fabrication, when those details are less prone to selection bias. By having less biased details (e.g., less detection bias) on how researchers go about fabricating data, it facilitates better contextualized simulations to investigate the efficacy of these statistical tools. After all, the results of these simulations are only as good as the set of data fabrication behaviors that are captured in the simulation.

In this report, we qualitatively assess the data fabrication strategies that practicing researchers used to fabricate data in one of our controlled studies (). Based on the transcripts of the interviews about how participants fabricated data (Hartgerink, Voelkel, Wicherts, & Assen, 2017), we apply qualitative methods to learn more about characteristics of the observed data fabrication strategies. Moreover, we combine the observed data fabrication strategies with results from various statistical tools to detect data fabrication [see also @], in order to assess whether certain data fabrication characteristics cause better or worse detection.

Methods

We used transcripts of 28 interviews with researchers who we previously asked to fabricate data in a controlled setting (available at Hartgerink et al., 2017). In these interviews, we asked participating researchers to answer questions separated into five sections. Section 1 pertained to general information about the researcher (e.g., frequent programs used). Section 2 inquired about the time and days spent on fabricating data (e.g., how many hours spent). Section 3 asked the researcher about their general framework with which they fabricated the data (e.g., what makes data look weird according to them). Section 4 focused on the specific steps taken to fabricate data (e.g., did they use a (pseudo-)random number generator). Section 5 was about the motivations of the researcher to participate in this study and their general assessment of their performance. All participating researchers consented to the public sharing of their transcripts.

To recapitulate, we previously asked these 28 researchers to fabricate raw data for a Stroop experiment (see Figure 1; **???**). In short, a Stroop experiment is typically a within-subjects experiment with two conditions measuring response times: (1) congruent (e.g., the word 'red' is presented in red) and (2) incongruent (e.g., the word 'red' is presented in green). We asked the participating researchers to fabricate response times for 25 participants, such that there was a statistically significant effect between conditions (i.e., a Stroop effect). Using these fabricated data (, we tested whether statistical methods could help separate fabricated data sets from (assumably) genuine datasets from Many Labs 3 (https://osf.io/n8xa7/; **???**).

In this paper, we take a two-pronged approach to evaluating the transcripts of these interviews. First, we provide qualitative summaries and reflections on each transcript. We include a description of the researcher (e.g., career stage, statistics knowledge) and how they fabricated the data according to the transcript. We also add what we considered noteworthy anecdotes from the interview. Secondly, we systematically compare what characteristics researchers applied to fabricate the data using Qualitative Comparative Analysis (QCA; Rihoux & Ragin, 2008). The first approach provides us with a more detailed but also less systematic picture of data fabrication, whereas the second approach provides us with a more general and more systematic picture

Stroop Task

Test of condition effect

t	df	p	Supported?
-20376.57	24	<.001	✓

	Congruent (milliseconds)			Incongruent (milliseconds)		
id	Mean	SD	Number of trials	Mean	SD	Number of trials
1	150	21	30	300	300	30
2	152	21	30	304	304	30
3	154	21	30	308	308	30
4	156	22	30	312	312	30
5	158	22	30	316	316	30
6	160	22	30	320	320	30
7	162	22	30	324	324	30
8	164	22	30	328	328	30
9	166	22	30	332	332	30
10	168	22	30	336	336	30
11	170	23	30	340	340	30
12	172	23	30	344	344	30
13	174	23	30	348	348	30
14	176	23	30	352	352	30
15	178	23	30	356	356	30
16	180	23	30	360	360	30
17	182	23	30	364	364	30
18	184	23	30	368	368	30
19	186	24	30	372	372	30
20	188	24	30	376	376	30
21	190	24	30	380	380	30
22	192	24	30	384	384	30
23	194	24	30	388	388	30
24	196	24	30	392	392	30
25	198	24	30	396	396	30

Figure 1: Example of a filled in template spreadsheet used in the fabrication process. Respondents fabricated data in the yellow cells and green cells, which were used to compute the results of the hypothesis test of the condition effect. If the fabricated data confirm the hypotheses, a checkmark appeared. This template is available at https://osf.io/2qrbs/.

of data fabrication.

Qualitative Comparative Analysis

In Qualitative Comparative Analysis (QCA; Rihoux & Ragin, 2008), qualitative information is deconstructed into characteristics and related to an outcome measure. In crisp set QCA, which we apply here, these characteristics are binary [e.g., present v absent;@rihoux2008]. Each unique combination of characteristics is regarded as a pattern and is used to assess necessary and sufficient conditions for the binary outcome measure to be present or absent. Using the coded characteristics for each unit of analysis (e.g., participants, group), we compile truth tables. Table xxxxx depicts a fictitious example of a truth table. For each unique combination of characteristics, the range of outcomes is inspected. The pattern `0-0-0` (first row) is observed ≥ 1 times and, in this sample, always leads to the absence of the outcome (vice versa for the last row). The pattern `0-0-1` (second row) is observed > 1 times and has conflicting (`C`) outcomes; both presence and absence occur with this pattern. The pattern `0-1-0` (third row) is not observed and therefore has no information about the outcome (i.e., a logical remainder; `?`). A truth table can subsequently be minimized to determine necessary and sufficient conditions for the outcome to be present or absent.

Table 1: Example of a truth table as used in crisp set Qualitative Comparative Analysis (csQCA). The outcome measure is the dependent variable, where the various patterns of the characteristics are used to determine under what conditions the outcome is observed. A `?` indicates that pattern was not observed and therefore the outcome is unknown; a `C` indicates that this pattern was observed >1, but that both outcomes occurred, creating a conflict in csQCA.

CHAR1	CHAR2	CHAR3	OUT
0	0	0	0
0	0	1	C
0	1	0	?
0	1	1	?
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

Based on the interview protocol (osf.io/xxxx), we identified five general data fabrication characteristics for our QCA. Each unique combination of data fabrication characteristics makes up a data fabrication strategy. We limited ourselves to five characteristics, considering that 2ˆn strategies would be possible. In other words, we balanced the number of transcripts (i.e., 28) to the number of unique data fabrication strategies possible (i.e., $2^4 = 16$; $2^5 = 32$; $2^6 = 64$). We coded whether (1) the participant prepared for the data fabrication (e.g., by reading literature on detecting data fabrication); (2) the participant used a (pseudo-)Random Number Generator (RNG) in fabricating the data; (3) the participant used assumably genuine Stroop data; (4) the participant duplicated or transformed data; (5) the participant checked the fabricated data for detectibility. The first author coded each of these five data fabrication characteristics for all of the 28 transcripts. Additionally, we coded ten participant characteristics (e.g., PhD attained, self-reported statistical knowledge; further described at osf.io/xxxx and available at osf.io/xxxx). We note that these data fabrication characteristics are inherently multiplicitous, hence, we do not know how much we will learn from the QCA.

As outcome measures, we included whether the researcher's fabricated data was detected as such, by taking the results from the three best statistical methods to detect data fabrication (**???**). In our original project, we included XX tests to detect data fabrication and included only the top three here, based on their Area

Under the Curve value. We could assess these AUCs based on (assumably) genuine data from the Many Labs 3 initiative (**???**). As a result, we included as outcome measures the results of the detection methods based on (1) ? (2) ? (3) ? We did not include the other methods, which consisted of (amongst others) X, X, and X.

We conducted separate csQCA analyses combining the coded data fabrication characteristics with the outcomes of these statistical detection methods. Usually, unique combinations in csQCA with conflicting outcomes are either omitted (comparable to listwise deletion) or additional characteristics are inductively added to resolve the conflicting outcomes (Rihoux & Ragin, 2008). Here, we omit conflicting outcomes for analysis and try to qualitatively assess potentially relevant characteristics. We do not run additional QCAs because adding additional characteristics quickly increases the state space of unique patterns to 128 (by just adding two) or beyond (512 by adding four). If we would add two characteristics it would result in maximum coverage of 22% by our participants; maximum coverage would be 5% if we add four characteristics.

We used the R package `QCA` (Dusa, 2007; R Core Team, 2017) to conduct these csQCAs. For each of the three statistical methods to detect data fabrication, we assessed both sufficient and necessary conditions for detection as well as going undetected. We minimized the truth tables using the enhanced Quine-McCluskey algorithm (**???**). Data are available at osf.io/xxxxx; analysis code is available at osf.io/xxxxx.

Results

Qualitative summaries

For each participant, we provide a qualitative summary of the interview with selected quotes. These summaries can be read separately from each other and provide detailed descriptions of our participants' fabrication process. Each summary is structured into at least five paragraphs: (1) how did the participant prepare for data fabrication, (2) how much time did the participant spend fabricating data, (3) what was the general framework with which the participant fabricated data, (4) what are the specifics to the data fabrication process, and (5) whether the participant checked the fabricated data for detectability. Generally, these are descriptions of blocks 2-5 of the original interviews (see Methods for a summary of the blocks). In order to remain neutral to the gender of the participant, we address the participant with their pseudonymized ID and with the pronoun 'they'. We share these summaries here and not in an appendix because this paper focuses on a qualitative evaluation of the data fabrication strategies applied. In Table xxxx we present an overview of the participant characteristics with hyperlinks to the raw transcript files (Hartgerink & Voelkel, 2017) and to the summaries provided in this manuscript (only available in the HTML version of this manuscript).

ID	Summary	PhD attained	Stroop experience	Software knowledge	Statistics knowledge [self-assessed, 1 (worst) - 10 (best)]	Confidence going undetected [self-assessed 1 (extremely insecure) - 10 (extremely confident)]
0jg	Link	No	No	[matlab, r]	8	
19e	Link	No	No	[r, python]	7	
1se	Link	Yes	Yes	[spss, r]	7	8

Participant `0jg`

Prior to fabricating the data, Participant `0jg` prepared by investigating existing literature on the Stroop task. The preparation by Participant `0jg` focused on assessing the scales and typical moments (e.g., standard deviations) of the response time measures. This preparation did not span more than half an hour and did not extend to investigating how previous cases of data fabrication were detected ("I thought it was a little bit cheating also to do it [check how data fabrication could be detected]").

Participant `0jg` spent approximately two hours on fabricating the data, on two separate days. The majority of the data fabrication took place on the first day and finished on the second day. Participant `0jg` also indicated that this was “not that much effort” from their part and that additional effort could have been put in by trying to investigate how we might go about trying to detect the fabricated data (“it was too much effort to do it”).

In order to make the data look less fabricated and more genuine, `0jg` indicated that the data should not follow the hypotheses or theory absolutely. In their own words, data would look more fabricated “if the difference between the conditions is equal for all participants - or between conditions” and “probably if it is [. . .] too normally distributed”. `0jg` also indicated that the test statistic for which they fabricated data affected the decisions made in the data fabrication process; “I noticed that in the [. . .] t-test that you did [. . .] the individual scores don’t matter [. . .] I didn’t pay too much attention to the size of the individual differences because I thought, [. . .] this doesn’t really matter anyway”.

`0jg` decided to fabricate the data by (1) taking a mean response time for one condition, (2) determine a condition effect, and (3) jitter the mean to get to the scores for individual fabricated participants. As such, “I had a mean of 545 with a difference between the conditions of 125 milliseconds.” Subsequently, `0jg` added noise by drawing from a normal distributions with different variances. For the variances, “I took a difference of 6 meaning that the congruent was 6 lower (so 67) than the incongruent [which was?] 73”. Effectively, for the congruent condition noise was added by drawing from $N\ (0, 67)$ and for the incongruent condition drawn from $N(0, 73)$.

Upon fabricating the data, the participant checked for (1) too large effect sizes, (2) minimal deviation of the preconceived standard deviations, and (3) odd histograms. As a result, the participant repeated the data fabrication several times, but “did not after each time inspect [the] data thoroughly. I just looked it at [sic] and thought well, just throw in a different one and go ahead and run it again.” These checks were off the cuff and not strictly defined, considering the participant noted that they “kind of had it in the back of my mind [to make it look like real data] all the time but no specific steps for it.”

Participant `19e`

Prior to fabricating the data, Participant `19e` indicated to have not done any conscious preparation. “I guess I should have. That is a good idea. But I didn’t, no. [. . .] I read a few meta-analyses on the Stroop effect. Hopefully, that gave me some insight. But no, I did not look at the literature on data fabrication itself.”

Participant `19e` spent approximately three hours on fabricating the data, on a single day. “I just put a block of [. . .] four or six hours in my calender and just devoted it to this thing.” Apparently, the participant took less time (three hours) than planned (four or six hours) and at the same time indicated “it was more difficult than I expected.” When asked how much effort they spent on fabricating the data, they indicated “It took some thinking as to what would constitute a good fabricated data set” and that they rated their own effort as 4 out of 7. In other words, it seems like they found fabricating the data more difficult than expected and did not put in much effort, which might explain why the time spent is shorter than blocked in their schedule.

In order to fabricate data and make them look genuine, `19e` set out to reuse a (supposedly) genuine data set. They indicated that they could “just look at [their] own old data [. . .] the means and standard deviations or maybe do even the calculations per subject [. . .] throw some noise over it or something.” But they figured they could use “a large data set and treat it as some kind of population and sample from that [. . .] to maybe avoid detection.” Reflecting upon the approach, “the first option [using own data] would be easier [to detect], I guess. [. . .] I think the approach that I took now [finding a large dataset as an artificial population to sample from] would be less easy to detect.”

Participant `19e` indicated several key points that would make data look more fabricated, including duplication and severe deviations in summary statistics. More specifically, they mentioned that “repetition of data [. . .] would be evidence for fabrication.” They also indicated to realize the data should be in accordance with what was previously published; “I think the best indicator for fabrication would be means or standard deviations that would not be similar to what has been found in earlier studies.” Beyond that, they considered “cross-correlations between [. . .] measures” and that “real data is not perfectly distributed.”

After searching on the Web for `stroop.csv`, Participant `19e` found a large dataset to use as a population to sample from. They admitted they “don’t know if it is real” and the dataset contained 121 subjects, which was considered sufficient. For each of those subjects, they computed the means and standard deviations for both the congruent and incongruent conditions across the individual trials. Subsequently, they looked at the distributions of the data (“that looked pretty normal”), the summary statistics of the mean and standard deviations, and the correlation between the means and standard deviations (although they did not specify how). Using these pieces of information, they “generated a new 121 sized data set with those characteristics.” From that generated dataset, they sampled 25 subjects once. In other words, the process involved finding a genuine dataset, investigate several summary statistics and their summary statistics, simulate data using that information, and then sampling a subset of that simulated data.

Upon fabricating the data, the participant checked for (1) odd histograms, (2) outliers, and (3) comparable correlations and summary statistics as the original dataset. They did not check the histograms systematically — “it all looked pretty ok-ish.” For outliers, they “checked for crazy outliers [. . .] maybe, there were response times of below 100 milliseconds and that would of course be not possible.” Additionally, the participant indicated that > “I compared the characteristics of my data - so, the mean of means and the standard deviation of means, mean of standard deviations, and standard deviation of standard deviations and all the cross-correlations between those measures - and I made sure that they were quite similar to in the real data set that I found.”

Participant `1se`

Prior to fabricating the data, Participant `1se` read one of their own papers on the Stroop effect. “I read one of our own papers on the Stroop effect [. . .] I needed [. . .] the means and standard deviations to have a starting point [for fabricating data].” Additionally, they recruited one of their postdocs to help out with the data fabrication. “I didn’t do this on my own. I recruited additional help [. . .] I don’t think we discussed it more than twice in total.” Participant `1se` was particular in the sense that they first actively rejected participation in the study. “I don’t want to participate” they said; “my first gut instinct was: I am not going to fabricate data. That is really bad.” After some back and forth about the goals and procedure of the study, they decided to participate after all. “I understand why it is necessary that we actually increase our ability to detect fabricated datasets that will be in the benefit - in the long-term benefit of science.”

Participant `1se` spent approximately six hours on fabricating the data, across four separate days. They indicated that they did not spend much effort, estimating it at a 2 on a scale from 1 (minimal effort) through 7 (a lot of effort). This somewhat conflicts their verbal assessment of the data fabrication process; “it is not difficult to fabricate data, unfortunately. But it is difficult probably to fabricate data that look like real data.” This could either be taken as that they did not try to fabricate data that looked real, that their recollection minimizes effort, or a variety of other factors. The first seems implausible, because the participant indicated that “if this [fabricating undetectable data] is the mission [. . .] we gonna [sic] beat them.”

Participant `1se` indicated that data look more fabricated when they fullfill three criteria: (1) too perfect normal distribution, (2) implausible absolute values, and (3) no outliers. More literally, they said data look fabricated when “they clearly come from a normal distribution” and “they are devoid of clear outliers”. For plausibility of absolute values, an example was given: “If the response latency would be 6 seconds, that would be very silly because the response latency is known to be - I don’t know from the top of my head - but 4-500 milliseconds for a congruent trial and [?] some 2-300 milliseconds longer.” These three criteria served as the basis for fabricating genuine-looking data. Criteria one and three seem somewhat contradictory, because they indicated that they wanted to fabricate a datset that was “at least semi-normal but not perfectly normal.” However, in a perfectly normal distribution there are by definition outliers, whereas in a semi-normal distribution there could be situations that there are fewer or no outliers.

“we deliberately introduced outliers” “We deliberately deviated from a normal distribution and our starting points always were realistic value for the Stroop effect.” “we simulated the data” “This is just a multivariate normal function from R which - as you know - uses a random generator” “we took into account the fact that the absolute latency has a relationship with the standard deviation” “If you are very slow then the standard deviation will be a little bit longer. So, there was a correlation between the mean and the standard deviation introduced." "we used the means from one of our own studies but we squared them to get skew and then we introduced the random fluctuating response latency - minimum response latency. And then we did some weird things just to deliberately delete things - deliberately make it not look like a standard normal distribution. We introduced some outliers at random - ourselves by hand. And then we did the weakest part. We modeled the within-person standard deviations as a non-central chi-squared distribution with a mean of the standard deviation of around 30% of the mean" "we rounded everything to milliseconds and then we wrote it into a SPSS file to hide the fact that we did it in R." "We simulated hundred subjects because we forgot that it had to be thirty. I think we did it maybe like 10 times or so. Randomly picking either congruent or incongruent. But, in all honesty, it was really done by hand and also not written down. So, this was haphazard. I couldn't reproduce it the next time." "then we transported it to SPSS and then from SPSS I had to save it as an Excel file and then copy" "No [use real data]. That was actually my first idea to simply permutate a real existing dataset. But we only used the means and the latency coming from a realistic study. But not the individual data from subjects."

"we simply used an algorith that should produce realistic data but we didn't check ourselves whether we could detect that they(?) were somehow fabricated" "We could have done a better job by looking into realistic standard deviations within subjects but we had no time to do that." "we were just too lazy to do that" "we didn't check the final results against the actual means and standard deviations that we sort of gave the simulator" [wrt plots] " we just looked at whether we didn't completely miss - just the very very eye balling - rough eye balling whether everything looked sort of ok-ish." "no formal checks if the value is like 2% outside of the original means we do it again until … That, we didn't do. We didn't iterate the process. Just did it one time." "We wanted to beat the system but not spend too much time." "We introduced outliers, we deliberately did not go for an all too normal distribution, and we stayed close to realistic values, namely those that are based on a real paper of our group." "We hope so [wrt difficult to detect as fabricated]. That was the aim."

csQCA

Author's note

All materials used in this project are available at https://github.com/chartgerink/2015ori-2 and are preserved at Zenodo. This project was funded by the Office of Research Integrity (ORI-).

References

Borsboom, D. (2013). Derailed: The rise and fall of diederik stapel. Retrieved from http://web.archive.org/web/20171221105716/http://www.psychologicalscience.org/observer/derailed-the-rise-and-fall-of-diederik-stapel

Dusa, A. (2007). User manual for the qca(GUI) package in r. *Journal of Business Research*, *60*(5), 576–586. http://doi.org/10.1016/j.jbusres.2007.01.002

Haldane, J. B. S. (1948). The faking of genetical results. *Eureka*, *6*, 21–28. Retrieved from http://wayback.archive.org/web/20170206144438/http://www.archim.org.uk/eureka/27/faking.html

Hartgerink, C. H. J., & Voelkel, J. V. (2017, July). Transcripts of 28 interviews with researchers who fabricated data for an experiment. http://doi.org/10.5281/zenodo.832491

Hartgerink, C. H. J., Voelkel, J. V., Wicherts, J. M., & Assen, M. A. van. (2017, July). Transcripts of 28 interviews with researchers who fabricated data for an experiment. http://doi.org/10.5281/zenodo.832490

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R

Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rihoux, B., & Ragin, C. C. (2008). *Configurational comparative methods: Qualitative comparative analysis (qca) and related techniques*. Thousand Oaks, CA: Sage Publications.

Stapel, D. (2012). *Ontsporing*. Prometheus Amsterdam.

(). Retrieved from https://github.com/chartgerink/2015ori-1