The European Political Science Association, 2018 *doi:10.1017/psrm.2018.45*

Online Appendix

Causal Modeling with Multi-Value and Fuzzy-Set Coincidence Analysis

MICHAEL BAUMGARTNER and MATHIAS AMBÜHL*

Appendix A: Background on Homogeneity, Model Ambiguities, Correctness

High consistency and coverage scores increase the reliability of the causal models output by CNA but do not guarantee their correctness. To get a clear understanding of the scope, inferential potential, and limitations of CNA, this appendix spells out what it means for the output of CNA to be correct and under what conditions CNA will certainly produce a correct output.

Very generally put, to say that CNA—or any other method—is a *correct* procedure of causal inference means that the causal conclusions it draws from data δ are true of the δ-generating causal structure Δ. This general characterization calls for two specifications. First, no method can be expected to systematically infer true models from deficient data. Whether data meet required quality standards depends on whether they faithfully reflect the causal structure that generated them. But since this structure is unknown in real-life discovery contexts, data quality cannot be assessed analytically but must be imposed by assumption (Cartwright 1989, 55-90). Even heuristics designed to ensure compliance with these assumptions, such as randomization and experimental control, cannot eliminate the risk of insufficient data quality. Accordingly, all procedures of causal inference come with a set of background assumptions, and are only guaranteed to produce correct results provided these assumptions are satisfied.[1]

While in most methodological traditions, the details of these background assumptions are thoroughly investigated and debated, the CCM literature has largely sidestepped this

*Appendix A draws on common work with Alrik Thiem, cf. Baumgartner and Thiem (2017b).

[1]For instance, regression analytic methods impose the *Gauss-Markov* assumptions (Gelman and Hill 2007, 45-47), and Bayesian network methods rely on the *Causal Markov* and *Faithfulness* assumptions (Spirtes, Glymour, and Scheines 2000, 29-31).

important issue so far. We cannot exhaustively fill this gap here (which would require a study in its own right), but still want to provide one background assumption—the configurational *homogeneity* assumption—which is sufficient to ensure the correctness of CCMs by ensuring that the analyzed data are not confounded (cf. Baumgartner 2009).[2] Generally put, configurational data are *confounded* iff unmeasured causes change between observed cases in such a way that variations in the outcomes appear to be due to the measured factors, whereas they are actually due to the changing unmeasured causes. Factors that can induce confounding are unmeasured causes of a scrutinized outcome Y that change the value of Y in a way that is not mediated through the measured factors in an analyzed factor frame **F**, i.e. causes of Y that are connected to Y on at least one causal path that does not go through the elements of **F**—so-called *off*-**F**-*path* causes of Y.[3] Changes in off-**F**-path causes of Y can bring about changes in Y that are erroneously ascribed to a measured factor that merely happens to co-vary with Y without being causally relevant to Y. Configurational data δ are not confounded if all off-**F**-path causes of Y remain constant across all cases in δ. Accordingly, an assumption that is sufficient to exclude confounding stipulates that δ are *homogenous* in the following sense:

Configurational Homogeneity (CH): Configurational data δ for an outcome Y over a factor frame **F** are homogenous iff every off-**F**-path cause of Y remains constant in all cases in δ.

Requiring δ to be homogenous in this sense amounts to a strong assumption that may be difficult to justify in observational studies. In fact, whenever the coverage of Boolean causal models is non-perfect, it follows that confounders are operative, meaning that **CH** is violated. A violation of **CH**, however, does not entail that causal inferences are impossible or that incorrect models will automatically be generated, it only follows that the correctness of resulting models is no longer guaranteed. Depending on how much risk a researcher is willing to take in a given discovery context, higher or lower degrees of **CH**-violations (e.g. visible in coverage scores) will induce her to abstain from a causal inference. On a par with background assumptions in other methodological frameworks, the function of **CH** is not to determine when causal inferences are possible but merely to *guarantee* the correctness of resulting models. If data δ are homogenous, it follows that all observed differences in the outcomes must be due to variations of the measured factors, which,

[2] We have to leave it to future research to determine whether the homogeneity assumption is also necessary for that purpose, or whether there exist alternative, possibly weaker assumptions that could likewise guarantee CNA's correctness. Moreover, note that data confounding is, of course, not the only data deficiency that can induce causal fallacies, errors of data collection (e.g. measurement error or selection bias) being another common type of data deficiency. For the purposes of this paper, we bracket errors of data collection by assuming that data have been faultlessly collected. Likewise, we do not consider misapplications of the method as a possible source of causal fallacies.

[3] This terminology is derived from Woodward's (2003, 59-60) notion of an *off-path variable*.

in turn, ensures that CNA cannot commit fallacies by ascribing the difference-making relations it uncovers to causal influences of the measured factors.

The second necessary specification of the rough characterization of the correctness criterion concerns the phenomenon of model ambiguities. There often exist multiple causal models that fit data equally well, to the effect that the data underdetermine their own causal modeling. Model ambiguities are a very common phenomenon in all methodological traditions (Simon 1954; Spirtes, Glymour, and Scheines 2000, 59-72; Eberhardt 2013; Baumgartner and Thiem 2017a).[4] Of course, CNA—on a par with any other method—cannot disambiguate what is empirically underdetermined. Rather, it must draw those and only those causal conclusions for which the data *de facto* contain evidence. In cases of empirical underdetermination it must, therefore, render transparent all data-fitting models (and leave the disambiguation up to the analyst). Multiple models in a CNA output are to be interpreted *disjunctively*, meaning that if, say, three models $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$ are returned, CNA determines that the data-generating structure has the form of $\mathbf{m}_1$ *or* that of $\mathbf{m}_2$ *or* that of $\mathbf{m}_3$. Such a disjunction is true iff at least one disjunct is true. Hence, in order for CNA to pass as a correct method of causal inference the data-generating structure must be truthfully reflected by at least one generated model.[5]

Overall, for CNA—or any other CCM—to be a correct method of causal inference it is required that at least one model inferred from homogenous data truthfully reflects the Boolean causal properties of the data-generating structure. More explicitly:

Configurational Correctness (CC): A configurational comparative method $\mathcal{P}$ is a correct procedure of causal inference iff, whenever $\mathcal{P}$ infers a set of models $\mathbf{M}$ from data δ which comply with **CH**, (at least) one model $\mathbf{m}_i \in \mathbf{M}$ satisfies the following four conditions:

(1) all values of exogenous factors contained in $\mathbf{m}_i$ are causally relevant for the corresponding outcome in the δ-generating structure Δ;

(2) if X_1 and X_2 are contained in two different disjuncts in $\mathbf{m}_i$, then X_1 and X_2 are located on two different causal paths in Δ;

[4] As shown by Baumgartner and Thiem (2017a), model ambiguities are much more frequent in configurational causal modeling than is typically acknowledged. In particular, applications of QCA are affected by a widespread practice of model-underreporting, one main reason being that the dominant QCA computer programs—as **fs/QCA** (Ragin and Davey 2016) or **Tosmana** (Cronqvist 2017)—regularly fail to uncover the whole model space, even for ideal data. While **QCA** (Duşa 2007) can avoid this problem if default parameter settings are appropriately tweaked, the only currently available QCA program that recovers the whole model space by default is **QCApro** (Thiem 2018).

[5] An analogous correctness benchmark is implemented in other methodological traditions. Spirtes, Glymour, and Scheines (2000, 81), for instance, require that a correct method returns a pattern of models (i.e. not an individual model) that represents the faithful indistinguishability class of data-fitting models, where a pattern is a disjunction (or class) of models. Similarly, Kalisch et al. (2012, 7), who require their procedures to only report the equivalence class of models in which the true model must lie.

(3) if X_1 and X_2 are contained in the same conjunct in $\mathbf{m}_i$, then X_1 and X_2 are part of the same complex cause in Δ;

(4) if X_1 and X_2 are two links of a causal chain in $\mathbf{m}_i$, then X_1 and X_2 are two links of a causal chain in Δ.

To a model $\mathbf{m}_i$ that truthfully reflects Δ by complying with conditions **CC**(1) to **CC**(4) we refer as a *correct model*.

We claim that CNA is a correct procedure in the sense defined by **CC** and provide substantive evidence for this in the main part of the paper. Two aspects of this claim deserve separate emphasis. First, that CNA is correct does not entail that it infers causal models from every data input. Data may be insufficient to warrant any causal inference. Whenever CNA abstains from an inference, it cannot commit a causal fallacy. By extension, correctness cannot be violated. Configurational causal modeling imposes very high quality standards on the processed data. If these standards are not met, a reliable CCM must refrain from drawing inferences. As detailed in the main part of the paper, CNA adopts a much more risk-averse approach in dealing with data deficiencies than QCA. While the latter does not impose a coverage threshold at all and often causally interprets minimally sufficient conditions that do not meet the consistency threshold, the former uses both consistency and coverage as *authoritative* model building criteria such that, if they are not met, CNA abstains from a causal inference. It is better not to draw a causal inference than to draw a hazardous one.

Second, that CNA is a correct method does not entail that it always *completely* uncovers the data-generating structure Δ. Real-life data tend to be fragmentary, meaning they do not contain all configurations that are empirically possible, that is, compatible with Δ.[6] Fragmentary data may not contain evidence for certain features of Δ, and no method can compensate for lacking evidence. Correctness merely demands that, if CNA outputs a set $\mathbf{M}$, then at least one model $\mathbf{m}_i \in \mathbf{M}$ be such that all causal properties represented by $\mathbf{m}_i$ truthfully reflect *some* causal properties of Δ. At the same time, if CNA is given exhaustive data featuring *all* empirically possible configurations, CNA should completely uncover Δ. That is, *completeness* is imposed as a conditional criterion: if CNA is given exhaustive data in compliance with **CH**, the Boolean causal properties represented by at least one model $\mathbf{m}_i \in \mathbf{M}$ truthfully reflect *all* Boolean causal properties of Δ.[7]

[6] *Data fragmentation*, as we use the term here, is related but not synonymous to *limited diversity*, a concept known from QCA (e.g. Ragin 2008, 147-148). QCA-processed data are said to be limitedly diverse iff they do not contain all *logically possible* configurations of the exogenous factors. CNA, by contrast, allows for the factors that are exogenous with respect to some ultimate outcome to be mutually causally dependent, in which case not all logically possible configurations are also empirically possible. Accordingly, we say that data are fragmentary iff they do not contain all *empirically possible* configurations.

[7] In Baumgartner (2009), an assumption of *empirical exhaustiveness* (PEX) is introduced to ensure that CNA-processed data is non-fragmentary and that Δ could be completely uncovered. We dispense with that assumption here. As a result, CNA will not always completely uncover Δ.

Since data fragmentation is ubiquitous in observational studies, procedures employed in this domain usually will only uncover a proper part of Δ. Still, if the data δ are fragmentary, CNA will uncover all those parts of Δ for which δ contain evidence, no fewer and no more. More specifically, although CNA is not unconditionally complete, it is unconditionally *informative* in the following sense: all and only those Boolean causal properties of Δ for which δ contain evidence are truthfully reflected by at least one model $\mathbf{m}_i \in \mathbf{M}$.

References

Baumgartner, Michael. 2009. "Inferring Causal Complexity." *Sociological Methods & Research* 38:71–101.

Baumgartner, Michael, and Alrik Thiem. 2017a. "Model Ambiguities in Configurational Comparative Research." *Sociological Methods & Research* 46 (4): 954–987.

———. 2017b. "Often Trusted But Never (Properly) Tested: Evaluating Qualitative Comparative Analysis." *Sociological Methods & Research* doi: 10.1177/0049124117701487.

Cartwright, Nancy. 1989. *Nature's Capacities and Their Measurement.* Oxford: Clarendon Press.

Cronqvist, Lasse. 2017. *Tosmana: Tool for Small-N Analysis [computer programme], Version 1.53.* Url: http://www.tosmana.net. Trier: University of Trier.

Duşa, Adrian. 2007. "User Manual for the QCA(GUI) Package in R." *Journal of Business Research* (URL) 60 (5): 576–586.

Eberhardt, Frederick. 2013. "Experimental Indistinguishability of Causal Structures." *Philosophy of Science* 80 (5): 684–696.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

Kalisch, M., M. Maechler, D. Colombo, M. H. Maathuis, and P. Buehlmann. 2012. "Causal Inference Using Graphical Models With the R Package *pcalg*." *Journal of Statistical Software* 47 (11): 1–26.

Ragin, Charles C. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond.* Chicago: University of Chicago Press.

Ragin, Charles C., and Sean Davey. 2016. *fs/QCA: Fuzzy-set/Qualitative Comparative Analysis, version 3.0 [computer program].* Irvine: University of California.

Simon, Herbert A. 1954. "Spurious Correlation: A Causal Interpretation." *Journal of the American Statistical Association* 49 (267): 467–479.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search.* 2nd ed. Cambridge: MIT Press.

Thiem, Alrik. 2018. *QCApro: Advanced Functionality for Performing and Evaluating Qualitative Comparative Analysis [computer program]. R Package Version 1.1-2.* URL: http://www.alrik-thiem.net/software/.

Woodward, James. 2003. *Making Things Happen. A Theory of Causal Explanation.* New York: Oxford University Press.

Appendix B: Additional Test Scores

Ratios of No Models Being Produced

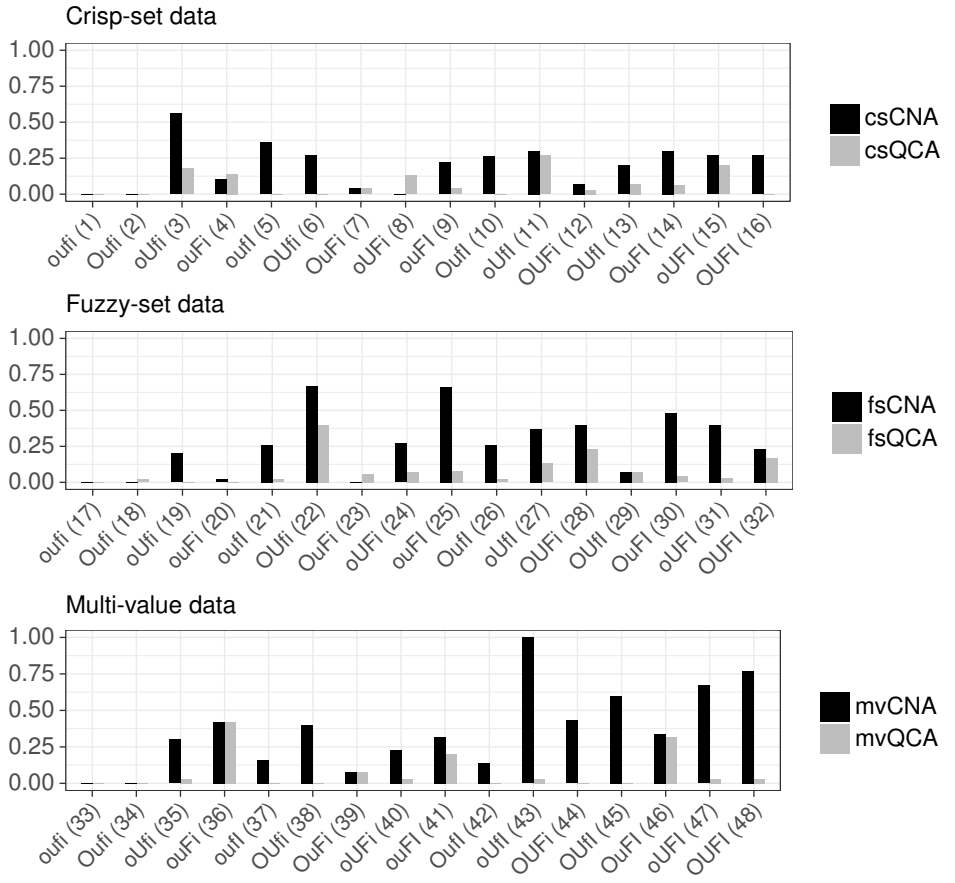

Figure 3. Ratios of trials in each test type in which no model is produced. The tests are numbered in correspondence with the replication script.

Ratios of Multiple Models Being Produced

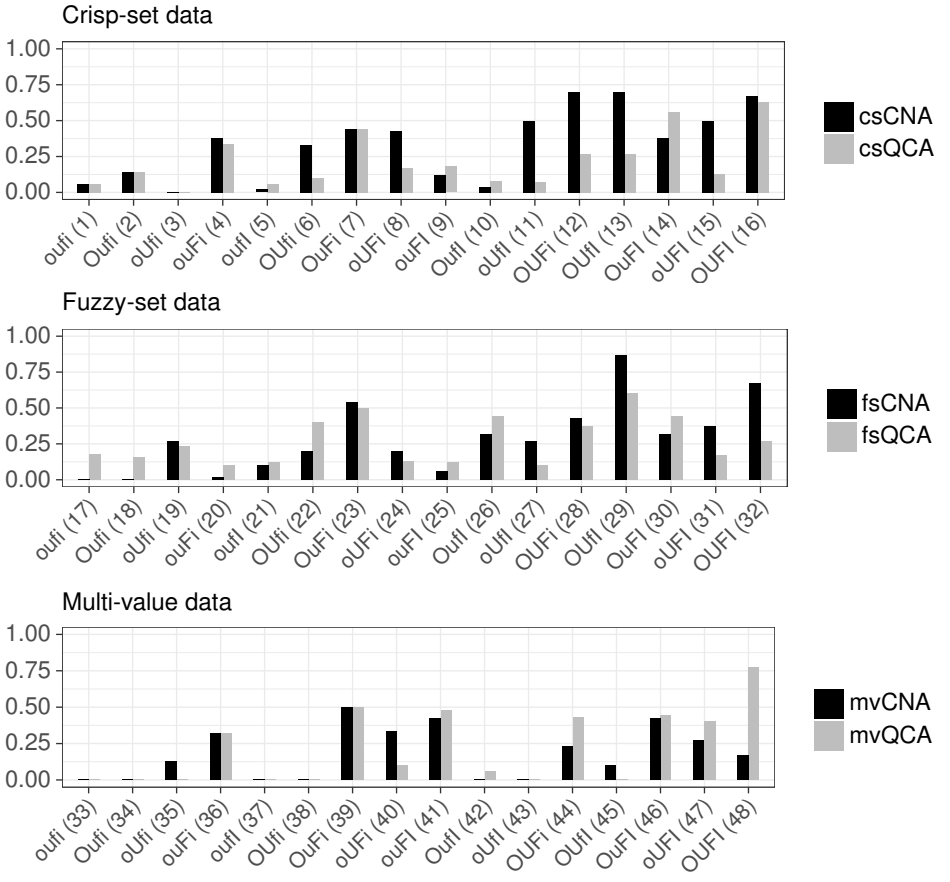

Figure 4. Ratios of trials in each test type in which more than one model is produced. The tests are numbered in correspondence with the replication script.

Ratios of Unique Models Being Produced

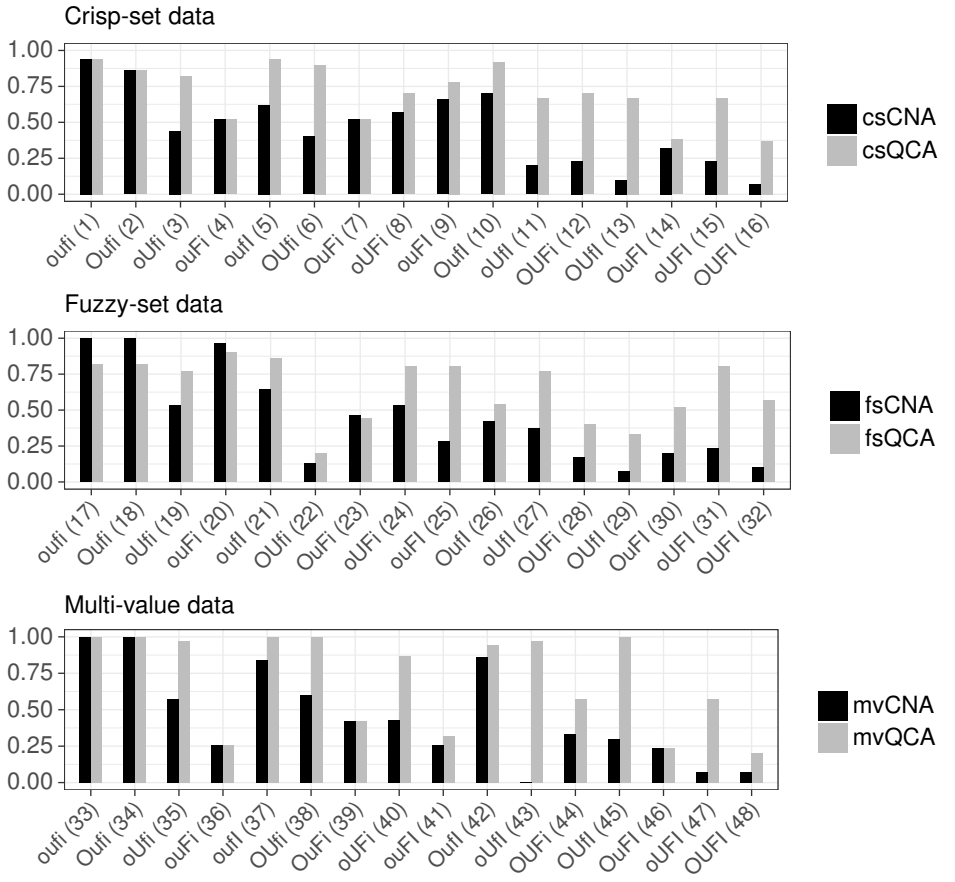

Figure 5. Ratios of trials in each test type in which one unique model is produced. The tests are numbered in correspondence with the replication script.

Ratios of Correctness Satisfaction by a Unique Model

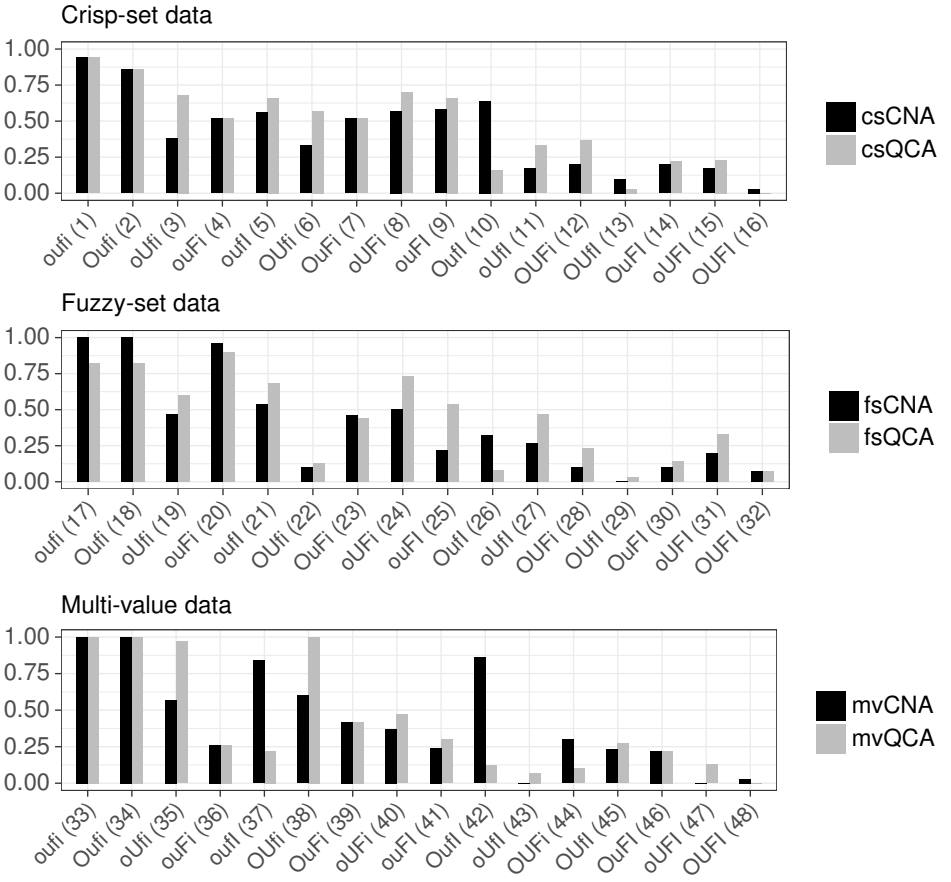

Figure 6. Ratios of trials in each test type in which correctness is satisfied by a unique model, i.e. such that exactly one model is issued which is correct. The tests are numbered in correspondence with the replication script.