Peer-review report of

Park, K., Waldorp, L. J., & Ryan, O. (2024). Discovering cyclic causal models in psychological research. *advances.in/psychology, 2*, e72425. https://doi.org/10.56296/aip00012

Round 1

Dear authors,

Reviewer 1 provided a detailed critique of your paper “Discovering Cyclical Causal Models in Psychological Research”.

After reviewing the comments and the paper, I believe the paper should be revised, and all aspects raised by the reviewer should be carefully addressed. I agree with the reviewer that the paper is very innovative and very well-written. Therefore, I encourage you to revise your paper with an eye towards 1) addressing the issues raised by the reviewer, 2) making sure to explain technical details/terminologies in a way that would be easier for both applied researchers and quantitative psychologists not familiarized with the field of Cyclical Causal Models, 3) using visualizations to describe/explain technical details whenever possible (the authors already did a great job presenting several Figures depicting different types of causal structures, and I encourage you to keep them all and maybe add more visualizations that can help readers understand the technical details of the field). I also strongly encourage you to blend your R code with the text in the empirical results, to help readers follow the analysis in a more tutorial-like manner. This strategy is highly effective in both improving the impact of the paper and helping applied researchers or quantitative psychologists new to the field of Cyclical Causal Models to learn how to use the techniques you’re applying/developing.

It is important to note that the reviewer's comments cover a wide range of concerns related to the paper's clarity, terminology, definitions, and interpretations. If the authors disagree with specific comments, they should provide clear explanations for their disagreements.

My editorial policy is to allow a free flow of idea exchange between authors and reviewers, and not to use peer-review as a gate-keeping process in which the reviewer has the final say.

In summary, Reviewer 1 offers a detailed critique of the paper and finds several major issues, including concerns about the definition of cyclic graphs, the distinction between DAGs and Bayesian networks, the interpretation of undirected graphs, and the handling of latent confounding. They also suggest clarifying the distinction between data-generating processes and hypotheses, improving the description of the CCD algorithm, and addressing redundancy in the writing style.

The reviewer raises minor issues related to terminology, the definition of d-separation, references, and the presentation of simulation results. They also suggest considering the results of the PC-algorithm in the empirical example and reducing detail in the discussion of cycles/chains. I particularly think that details in the discussion of cycles/chains are informative, but you can use your own judgment if you will make it more concise or not.

In summary, the reviewer's feedback highlights several areas where the manuscript can be improved regarding clarity, accuracy, and presentation, particularly in addressing key conceptual issues related to causal discovery and graph modeling.

Best wishes,

Hudson Golino

Reviewer 1

As an enthusiast of causal discovery and its applications to psychology, reading this study was quite exciting. The aim of the study was to provide an accessible introduction to the basics of cyclic causal discovery for empirical researchers. Overall, I believe it is both well-written and well-cited and it was able to achieve its general aim. However, I identified some limitations that I believe need further considerations before the manuscript can be considered for publication:

Major issues:

From a more abstract/philosophical perspective, cyclic graphs are impossibilities with cross-sectional data. I believe this point was made by Pearl in his "Causality" book and I know that Glymour and Spirtes also defend this idea somewhere. But from a deterministic and realist/localist perspective, it is not possible that a cause is an effect at the same time that an effect is a cause. Causes have temporal precedence over effects in a physical word, with the only agreeably known exceptions being some quantum phenomena. I believe it is important to stress this point because network models have been wrongly defined in the psychological literature as models of "dynamic" and "mutually reinforcing" aspects of the studied construct, but this definition is of course only valid when the data is longitudinal. Therefore, I believe the authors should make some considerations about this in the introduction (even if they disagree with it) and also make it clearer that the type of "cycles" that they are talking in here are relations that are not causally identifiable within the definition of DAGS.

A DAG is not a Bayesian network. A DAG is a pair (V, E) with no directed cycles. In a more insightful definition, a directed graph is a DAG iff it can be topologically ordered (i.e., one can create a total order of the vertices with the relations based on the directions of the edges). A Bayesian network is a DAG where V are variables and E represent directed (causal) dependence relations. In this sense, a DAG is not a "causal" graph, but an isomorphism of the causal dependencies represented by the Bayesian structure of the data. This discussion is probably too technical to include in this paper. However, because probabilistic graphical modeling is usually misrepresented because of some misconceptions, especially in psychology, I believe the authors could try to make clearer distinctions about the abstract mathematical tools used to talk about the causal relations and the statistical methods implemented to interpret these causal relations. In another example, the PMRF is not the same as a GGM, despite of the fact that you can operationalize a PMRF using a GGM. Therefore, I believe the authors should try to make it clearer (and they have done this in some parts of the paper) that the graph theory, the causality theory, and the

statistical/computational tools discussed in the paper, despite being all part of a larger discussion about causal modeling, are not always the same thing.

The authors used versions of this argument in several sections of the paper: "causal discovery methods specifically designed for this task are likely to outperform statistical network models in learning the underlying causal structure, indicating that network models may not be desirable tools for discovering causal relationships". But this argument relies on a misconception about how to interpret the edges of undirected graphs. Because the directions are not defined, undirected edges can equally represent marginal or conditional dependencies. However, many researchers wrongly interpret the edges only as marginal dependencies. In this sense, more generally, undirected graphs can be defined as equivalence classes of directed graphs, with the size of the set of equivalencies determined by the additional causal assumptions one is willing to make or, in the case of experimental data, one can actually make based on the design of the study. For instance, a strongly connected undirected graph with three variables X, Y, and Z can be the result of basically any data generating process. However, if one assumes the Causal Markov condition, faithfulness, and causal sufficiency, and if the variables X and Y are marginally independent, then the DAG {X -> Z, Y -> Z} is the only element of the equivalence class determined by the undirected (moral) graph. This relation between undirected and directed graphs means that, therefore, network models are not "worse" in learning the underlying causal structure, they just don't make the additional assumptions that are necessary to direct the marginal dependencies and to remove the conditional dependencies. In fact, some causal discovery algorithms (e.g., PC-algorithm and GES) use network models as their initial estimation step as the resulting directed graph will necessarily be an element (or a more restricted equivalence class, such as a CPDAG) of the equivalence class represented by the undirected graph. Of course, I am being repetitive and some of these questions were already explained by the authors. But I suggest that some parts of the text should be revised to avoid misleading conclusions that readers can get to.

Versions of this statement have been repeated in some parts of the paper: "In general, constraint-based techniques are unable to uniquely identify the underlying causal graph, but instead return a set of causal graphs that imply the same statistical independence relations". However, this limitation is not unique to constraint-based algorithms, but actually a limitation imposed by the laws of causality (as discussed by Pearl) to any causal discovery algorithm applied to correlational data. The authors do explicitly state this in later sections of the paper, but sometimes a statement equivalent to the one in the beginning of this paragraph is not properly explained, which may lead to misleading conclusions to some readers.

The authors have used this argument in some parts of the text: "the main advantage of the other two algorithms—FCI and CCI—is their ability to handle the presence of latent confounding, which commonly occurs in psychological research in practice". I believe the most important aspect of FCI and CCI (and of causal discovery algorithms that assume the presence of latent common causes, in general) is not that confounding is possible (or even "common" as it may be in psychology), but that, in fact, without the assumption of sufficiency, it is impossible to say that the observed relation between

two variables is a direct causal relation. In this sense, the authors can even improve the impact of their presentation by emphasizing that the generality allowed by causal discovery algorithms that make less assumptions may be more interesting to most researchers, even if they are not interested in DCGs.

I believe the authors should make the distinction about the data generating process (DGP) and the hypothesis about the DGP clearer. For instance, when describing Figure 10, I think many readers will not readily understand that what the authors meant is that the data can be originated by two different DGPs that will result in the same observed set of (in)dependencies and, therefore, one cannot be sure about what the true DGP was, only what the equivalence class is. There are some other parts of the text that I believe the readers would be benefited by the more explicit distinction about DGP and the hypothesis made about the DGP after causal discovery.

I believe the description of the CCD algorithm is necessary, but maybe it is a bit cumbersome. I don't have a satisfactory suggestion, but maybe have both the text and the diagrams in a figure? As the authors suggest that only the interested reader should spend more time with this information, maybe this is enough to summarize the information without removing most of it and putting in an appendix or supplemental material. Then, in the text they can have only a summary, as they did for the other algorithms.

The meta-commentary writing style of the authors is cumbersome sometimes. In some parts of the text, like when doing the summaries of larger sections, it is quite helpful. But when it is used in the first sentence of a new section, or in the end of a small section, it is a bit repetitive.

I believe it is redundant to talk about a "latent confounder", as "confounding" is defined in terms of the data generating process, which, by definition, is not observed. Also, if a confounder is observed, it just becomes a covariate or control.

There is one major problem with the simulation study. Because FCI and CCI are causal discovery algorithms that do not assume sufficiency, they will most likely than not fail to recover some arrows when the data generating process (DGP) does not include a confounder. On the other hand, CCD will never be able to identify that the relation between two variables may be due to a confounder. In this sense, CCD is not comparable to FCI and CCI because they actually have different objectives. I think it is even more evident if one were to include the PC-algorithm in the simulation. We know that it will necessarily fail in the proposed scenarios as it can only handle (CP)DAGs without selection bias and without confounders. CCD should necessarily fail in some scenarios and FCI and CCI should necessarily fail in other scenarios, but this didn't happen. I believe it happened because the indices are not adequate to reflect the performance of the algorithms. Maybe the authors can solve this problem by calculating statistics based on comparison of clusters, such as Variation of information or normalized mutual information or Rand index, as alternative measures to the SHD

on how different the true ancestral graphs are to the estimated ancestral graphs. These indices would also allow the authors to more clearly identify which algorithm better recovers the cycles.

Minor issues:

On page 2: The authors state that "In the field of causal discovery, using patterns of statistical (in)dependence estimated from observational data to infer causal structures is known as constraint-based causal discovery". However, this statement is incorrect. Constraint-based algorithms are one of the possible ways of using patterns of statistical (in)dependence estimated from observational data to try to discover causal relations in the data. Score-based, permutation-based and hybrid algorithms are other approaches for causal discovery that do not dependent on constraint-based methods. The authors also state that "invariance-based algorithms—another type of cyclic causal discovery methods". But there are invariance-based algorithms for acyclic causal models.

On page 2: "For example, in the context of structural equation modeling, it is well known that models are usually identified as long as they are acyclic, whereas this is not necessarily true for cyclic models". This sentence is a bit confusing: cyclic models are not identified when they are acyclic?

On page 2: "as they may be produced by unwittingly conditioning on common effects or by failing to account for unobserved confounders" the same may be true for cyclic graphs, but I believe the text is written in a way that the reader may be misled to believe that this does not happens for cyclic graphs.

On page 3: Formally, a graph is not a diagram, but a pair. Because the authors are using the formal notation to define what is a graph, maybe it would be more adequate to also use the formal terminology accordingly.

Figure 1: Maybe the authors should use directed graphs that come from the same equivalence class of moral graphs? For instance, the DAG {X -> Y, Y -> Z, X -> Z} has the same moral graph as the DCG {X -> Y, Y -> Z, Z -> X}. The example provided by the authors can be confusing, as the readers may be left under the impression that a cycle is present when one is not certain of the direction of the relation between two variables. However, cycles are related to the general structure of the graph. The original Figure 1 can still be used later, as the new Figure 2, when the authors discuss the d-separation in cyclic graphs.

On page 4: the authors should formally (and intuitively) define what the d-separation criterion is. The examples in this part of the text are more confusing than clarifying without a proper definition.

On page 6: "Together with the global Markov property, faithfulness enables us to make inferences about causal relationships represented in graphs by testing for the statistical independence among variables". If the causal discovery algorithm does not include unobserved causes, the assumption of causal sufficiency is also necessary.

On page 9: the original reference for FCI is "Spirtes, P., Meek, C., & Richardson, T. (1995, August). Causal inference in the presence of latent variables and selection bias. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (pp. 499-506)." I think CCI has also another original reference, but I can't find it right now.

On page 20: "However, the PAG output by CCD is by far the most informative, as it contains no circle endpoints, thus representing the smallest equivalent set of graphs." wouldn't the most general equivalence class be more informative?

Figure 11: these are exactly the DGPs for the simulation or just an example? I believe it is not sufficiently clear.

On page 29: "Therefore, they penalize an algorithm for making incorrect predictions but not for being conservative". I think the authors should consider "being conservative" also as making incorrect predictions and include this in the calculation of precision and recall. If I understood correctly, the authors already guaranteed that the directions of at least some of the edges are causally identifiable to begin with. Therefore, not being able to identify these directions should be counted as an error.

I believe the sensitivity analyses could be in an appendix and that the "intermediate summary" section is not necessary.

Maybe it would be interesting to include the results of the PC-algorithm in the empirical example as well.

I believe the authors don't need to describe the cycles/chains with so much detail in the text. This makes it difficult to follow the text.

I believe the second paragraph on page 38 to be unnecessary. It is redundant with other things that were said before and are stated again in the discussion. It would be more appropriate if the authors included some theoretical implications regarding the results they have found. I understand that they used the data only as an example. But it should also be clearer how one can make inferences from this type of analysis.

Round 2

Dear Dr. Park and colleagues,

Based on the authors' thorough responses to the reviewer comments and the revisions made to the manuscript, I recommend accepting this paper for publication.

You have carefully addressed each of the major and minor points raised by the reviewer. You have also clarified key concepts and terminology throughout the paper, such as the distinction between DAGs, Bayesian networks, and PMRFs. The revised introduction now provides important context about how cyclic causal models can be interpreted as equilibrium models fit to cross-sectional data.

You have thoughtfully engaged with the reviewer's critique of the simulation study. While acknowledging that some results were unexpected, such as CCD performing better than anticipated with latent confounders, they have added useful discussion and appendices exploring these findings in depth. The use of multiple performance metrics provides a balanced assessment, in my opinion.

The empirical example is clearer and more impactful with the addition of theoretical implications drawn from the CCD output. Relocating some technical details to the appendices has improved readability.

Although a few minor suggestions were not implemented, such as including the PC algorithm, you have provided reasonable justifications for your choices.

Overall, the revised manuscript is a rigorous and accessible introduction to cyclic causal discovery for empirical researchers. I deeply appreciate your diligent response to the constructive feedback provided by the reviewer. In my view, it has significantly strengthened the paper.

I believe your paper will make a valuable contribution to the field and recommend it be accepted in its current form. Now the paper will be sent for production, and you can expect the journal to contact you in case there’s any copyediting modifications necessary.

Best wishes,

Hudson Golino

Reviewer 1

I am glad to see that the authors have taken my suggestions/comments into account when revising their manuscript. I believe the text has improved considerably, and only some minor issues and one main issue remain. Despite of the fact that some of my comments may have been not as clear as I expected, I believe most issues were dealt with in an appropriate manner. Below are my comments regarding the current issues:

Major issue: A variation of this sentence shows up in some parts of the manuscript: “which algorithm is most suitable for situations likely to arise in psychological research”. I will insist that this is problematic, as the readers may be under the impression that the results in the simulation study reflect what is the “best method” for causal discovery. But because the causal assumptions are not testable in the dataset we use for causal discovery (e.g., if we only have X1 to X4 from Figure 10, we cannot know which method is correct), we can only know which method performs better under

a set of causal assumptions about the DGP. Overall, I think the major issue is that: (I) it is still not very clear that the causal discovery methods presented in this study will provide a range of causal hypothesis about the data that also include the DAGs provided by more traditional procedures; and (II) that the simulation is mostly illustrative rather than generalizable (e.g., if you included DGPs with structures made only of v-structures, based on the metrics, PC-algorithm would probably perform better).

Page 2: "However, it has been shown that network models are likely to perform poorly as causal discovery tools; relations in the network may not reflect the direct causal effects that researchers aim to discover, as they may be produced by, amongst other inferential issues, unwittingly conditioning on common effects". My only issue with this sentence (and others still present in the text) is that one should never interpret conditional dependencies as direct causal effects, at least not with additional causal assumptions that are not part of undirected networks. Therefore, maybe the issue is more on the side of researchers (misusing the method due to misconceptions) than on the method. But I am okay to agree to disagree.

Page 2: "In part, this is due to the conceptual and practical difficulties in fitting and interpreting cyclic causal models." Is "fitting" the most adequate term? It is used in other parts of the paper as well, but I believe this may create the expectation in the readers that the authors will teach how to estimate regression/correlation parameters, rather than causal discovery.

Page 3: "From this perspective, cyclic causal relations should be interpreted as a kind of coarse-grained or time-averaged representation of (reciprocal) causal relations between processes that evolve over time; for a detailed treatment of cyclic equilibrium causal models in the context of psychological modeling, we refer readers to Ryan and Dablander (2022)." I don't remember this being in the first version of the manuscript, but I really liked it! Together with the previous sentence, it properly aids the reader on correctly interpreting what a cyclic causal graph is.

Page 4: “These models assume particular distributions for the variables involved, and, in the case of the GGM, assume that conditional (in)dependence relations can be captured by linear dependence parameters such as the partial correlation.” The sentence starts with saying that the model assumes a distribution, but ends talking about dependencies. I understand that variables that are linear combinations of other variables follow a multivariate normal, but maybe rephrase this sentence to make this relation clearer.

Page 4: “Formally, two variables A and B are said to be d-separated given C if and only if all paths between A and B are blocked when conditioning on C”. Maybe explain what does it mean for a path to be blocked.

Page 5: “Two spurious edges are induced in the PMRF”. I believe some readers may be confused by the meaning of “spurious” in this and similar sentences.

Page 6: “Despite this limitation, in practice, PMRFs have often been interpreted as a causal skeleton—the undirected version of a causal graph”. Maybe “been incorrectly interpreted”? After all, the skeleton is completely different from a moral graph of a DAG.

Page 6: “However, these models are prone to suboptimal performance, as the equivalence class they identify is likely much larger than that of custom-built causal

discovery methods". I think this is a bit confusing. First, maybe depending on the audience, the authors should define what a "equivalence class" is. Second, do they mean here that the equivalence class of a (CP)DAG is larger than that of a PAG? Wouldn't that be the reverse, as a MAG represents several possible DAGs, and a PAG represents several possible MAGs?

Page 12: "PAG, which represents the common features of equivalent directed cyclic graphs". And also of equivalent DAGs.

Page 12: "however, three different types of edge-endpoints" but then the authors include a * endpoint in the description.

Page 13: "For example, from the PAG shown in Figure 4, we can read off the following" but the notation in Figure 4 is different from the one in the text.

Page 24: "All simulations are performed using R software version 4.2.3". What packages were used, if any, and for what purpose? If no packages were used, maybe make it explicit that all the functions were made by the authors.

Page 34: "To obtain a sparse network, we used the graphical lasso (glasso) method to regularize partial correlations". There are some studies showing that glasso is highly biased and some other regularization methods work better (see, e.g., the methods in the GGMncv R package).

Page 34: I agree to disagree on adding the PC-algorithm to the simulation study, but why not including it to the empirical analysis? I believe this could be helpful to show the different hypothesis one can draw from the methods, especially because there are many sentences on the manuscript discussing how learning DAGs can be restrictive.

I would like to restate my view that this is a very interesting study and probably the most intriguing paper I've read on causal discovery in a psychology journal.