




An Introduction to Complementary Explanation

Joeri van Hugten ¹

This paper introduces the practice of complementary explanation; the practice of taking a published result and writing a focused paper that rigorously and systematically describes the implications for a theory that would be rejected by those results. Such spotlighting of a rejected theory counteracts the common alignment between theory and result in published work.

Keywords *philosophy of science, falsification, publication bias, complementary explanation*

The reliability of the social sciences is threatened by underreporting. Underreporting refers to reported evidence not reflecting all collected evidence. This is concerning if reported evidence is a systematically disproportionate subset of collected evidence. Currently, this is the case with reported evidence being severely biased toward evidence that supports theories. For example, a large-scale replication of 100 psychology experiments replicated only 36 out of 97 significant results (Open Science Collaboration, 2015). In industrial organizational psychology, hypotheses in journal articles are 73% supported and 12% rejected, while hypotheses in dissertations (which should be more representative of all collected evidence) are only 33% supported and 42% rejected (Mazzola & Deuling, 2013; van Hugten & van Witteloostuijn, 2021).

Underreporting is caused by underreporting practices such as hypothesizing after the results are known (HARKing) and not writing up all conducted tests. This leads to bias because especially results that do not support a paper's theory tend to be the ones not written up, and hypotheses made after the results are known tend to be ones that are in line with those results. Underreporting practices are prevalent. Measuring socially undesirable behavior is difficult, but best efforts suggest that 91% of academics know faculty who engaged in HARKing in the past year, 77% knows faculty who selected data that would support their hypothesis and withheld the rest (Bedeian et al., 2010; Rubin, 2017).

This paper proposes a practice to chal-

lenge theories and counteract underreporting. That practice is based on the falsificationist hypothetico-deductive philosophy (van Witteloostuijn, 2016), because it opposes the beliefs underlying underreporting practices.

Underreporting practices as a neglect of falsification

Besides psychological factors (e.g., confirmation bias) and sociological factors (e.g., not undermining your colleague's theories), beliefs about what is important also underlie underreporting practices. In this section, I speculate about underlying beliefs for three aspects of underreporting practices, as well as a falsificationist principle that speaks to that belief. The posited beliefs are overlapping, and it turns out that falsificationist principles form a coherent opposition to those beliefs.

In the theory and hypothesis sections, why do HARKed hypotheses tend to be in line with the result? My personal intuition is that researchers understand that, at the level of the research program, theories aim to explain phenomena, but that this gets mistakenly transferred to believing that also hypotheses aim to explain results, at the level of the individual study. Given that aim, it follows that hypotheses that are not in line with the result are not useful (Johns, 2019). By contrast, falsificationist principles maintain that hypotheses aim to challenge theory. For that aim, also hypotheses that are not in line with the result can be valuable (as long as they are tightly connected to a theory).

¹Vrije Universiteit Amsterdam

Received August 29, 2021
Accepted May 12, 2022
Published August 15, 2022

Correspondence
Vrije Universiteit Amsterdam
j.g.w.j.van.hugten@vu.nl

License

This article is licensed under the **Creative Commons Attribution 4.0 (CC-BY 4.0)** license, which allows you to copy, redistribute, transform and build upon the article and all its contents in any medium or format for any purpose, provided that appropriate credit is given.

© van Hugten 2022



Table 1 Possible beliefs underlying underreporting practices.

<i>Underreporting practice</i>	<i>Possible underlying belief</i>	<i>Related falsificationist principle</i>
HARKed hypothesis tend to be in line with the result	Theories aim to explain the world, therefore hypotheses aim to explain the result	Hypotheses aim to challenge theory
Unreported results tend to be against the theory	Supported hypotheses give more valuable knowledge	Rejected hypotheses give more valuable knowledge
Discussion section focus on explaining the result	Explaining the result is the goal of a paper	Challenging the theory is the goal of a paper

Table 2 Complementary explanation in relation to existing practices.

	<i>Focus on before the result is known</i>	<i>Focus on after the result is known</i>
<i>Focus on the result</i>	Dominant practice	Abduction, CHarking, Harking, RHarking, Tharking
<i>Focus on the rejected hypothesis</i>	Counterargument, competing hypothesis, meaningful baseline, theory-driven null-hypothesis	Complementary explanation, SHarking

In the results section, why are results against hypothesis the underreported kinds of result? A possible underlying belief is that results that support hypotheses grant more valuable knowledge. In direct contrast, a key principle of falsification is that results against hypotheses give more valuable knowledge. For instance, seminal falsificationist Karl Popper argues that we learn more from results against hypotheses. Broadly speaking, the argument is that a result that supports a hypothesis does not imply that the theory is true, because that is affirming the consequent. Specifically, such an inference would go: ‘if theory T is true, then data D should be observed’ (i.e., the hypothesis), ‘data D is observed’ (i.e., the result is in line with the hypothesis), ‘Therefore, theory T is true’. This is a logical fallacy because there may be alternative explanations for data D. Therefore, researchers try hard to rule out such alternative explanations by using random assignment, control variables, or more advanced statistical techniques. By contrast, a rejected hypothesis does imply that at least one premise in the theory or operationalization is false, because it is

denying the consequent which is a valid form of argument (even if alternative explanations were not excluded). Less extremely, (Davis, 1971) influentially argues that results that go against our expectations are more interesting.

In the discussion section, if a rejected hypothesis is reported, why is the expectation that authors explain the result? I speculate that the underlying belief may be that explaining data is a more important goal than improving theory. By contrast, falsificationist principles hold improving theory as the main goal. Therefore, those principles suggest that discussion sections build on the result to contribute contingencies that make the theory less simple or generalizable, and as a result, more accurate (e.g., Cross, 1982; Lakatos, 1970). Contributing contingencies can also happen in the process of explaining a result. However, the distinction is especially clear when discussion sections bring in a completely different theory that does fit the result. The distinction also becomes clearer if one imagines a more extreme alternate world in which discussion sections purposefully attempt to bring in additional theories that are opposite to the result. By contrast, current practice is that no further discussion is needed once the result is explained.

Overall, the argument is not that following the principles of falsification leads to more ethical research; it probably only affects the type of results that are underreported, not the extent of underreporting. That is, if researchers believed that rejected hypotheses lead to more valuable knowledge, then underreporting might start tending toward underreporting results that support hypotheses. Currently, the tendency is to underreport rejected hypotheses, so a practice based on principles of falsification can help bring balance.

A proposed counteracting practice: complementary explanation

Because of the opposition to falsificationist principles in the aspects of underreporting practices, I propose that a practice that thoroughly applies those falsificationist principles can counteract underreporting practices. Specifically, I propose complementary explanation (CE).

The term ‘complementary explanation’ is a variation on the term ‘alternative explanation’.

Table 3 CE Steps

Steps	Notes
0. Find a result	While reading, one might stumble upon a published finding that is striking if interpreted from the perspective of a different theory. The finding might even be merely a control variable for the original paper. Results with strong measures and research designs are ideal so that the result being opposite to a CE is clearly attributable to the theory.
1. Develop a CE for that result	What collection of premises suggest the opposite of the result? Premises that are straightforward and commonly held associations of concepts are ideal. That collection becomes the CE. If the result is opposite to the original hypothesis, then the original hypothesis development is a CE.
2. Identify a premise in that CE to challenge	By design, the CE is not in line with the result. So, at least one of its premises must be too simple.
3. Suggest a complication for that challenged premise	What would be one way in which we could complicate the challenged premise?
4. Evaluate that complication's effect on accuracy.	Does the complication increase accuracy? Is the complication plausible?
5. Iterate over steps 3-4. When out of ideas, compare complications.	The most simple and generalizable complication that can accurately predict the result is the ideal.
6. Iterate over steps 2-3-4-5. When out of ideas, compare challenges.	The less paper-specific the challenged premise, the greater the theoretical contribution.
7. Specify the contribution	Concisely and concretely describe the new insight. E.g., 'Premise 1 should be replaced by premise 1*' or 'Premise 1 is moderated by M'.

An alternative explanation is an explanation for a result and an alternative to the hypothesis development (assuming that the result was in line with that hypothesis). Alternative explanations are the main threat that Popper aimed to avoid. By contrast, a complementary explanation (CE – countable) is an explanation for the opposite of a result, so it is a logical complement to the hypothesis development (assuming that the result was in line with that hypothesis). For example, if a quantitative study finds a positive coefficient, a CE for that result is a set of

arguments that imply a negative coefficient. Similarly, for a qualitative study's causal story between high X and high Y, a theory's implication of a negative relation is a CE. If a study's result is inconsistent with its hypothesis, then the original hypothesis development is a CE. Even if a study does not have a hypothesis for a particular relation, an explanation of the opposite of its result is a CE. One result can have multiple CEs.

To appreciate CE's unique focus, Table 2 positions CE in the context of a comprehensive list of similar existing practices. CE is similar to counterarguments, competing hypotheses, meaningful baselines, or theory-driven null hypotheses (e.g., Schwab & Starbuck, 2012). The difference is that CE is to be used after the result is known. Even more extremely, CE can be done after publication by someone who was not the original author.

CE is like HARKing and spinoffs like Tharking (i.e., transparently hypothesizing after the results are known (Hollenbeck & Wright, 2017; Rubin, 2017) and abduction (Locke et al., 2008; Schwab & Starbuck, 2017) in that all those practices happen after a result is found. However, the difference is that those practices aim to explain a result (although RHarking is, in principal, also open to rejected hypotheses (Rubin, 2017)). For example, abduction would never involve explaining the opposite of the result. In other words, hypotheses made after the results are known tend to be ones that are in line with those results. But they need not be that way. CE is like transparently making a hypothesis after the result is known, that is opposite to that result. That shift in focus counteracts the threat of HARKing to research reliability. Finally, CE is like SHarking (suppressing hypotheses after the results are known); the most threatening form of HARKing (Rubin, 2017), except that SHarking focuses on suppressing rejected hypotheses while CE adds exactly such hypotheses.

CE Steps

The steps to interpret supportive results seem clear: e.g., 1) $p < 0.05$, 2) hypothesis supported, and 3) more confidence in the theory (but see Wasserstein et al. (2019) for how it is not that simple). By contrast, the application of falsification is impeded by a lack of such clear steps. CE

is a way to codify falsificationist interpretation steps. Table 3 summarizes these steps.

A crucial step in falsification is that a rejected hypothesis implies that at least one premise in its explanation is false. But, it is undetermined exactly which one is false (Hines, 1988; Lakatos, 1970; Sørberg, 2005). That underdetermination can make falsification seem infeasible in practice. CE tackles this issue by evaluating theories based on a combination of their accuracy, simplicity, and generalizability (Weick, 1999). If a result is against a theory, that means the theory has low accuracy. Then, we can trade off generalizability or simplicity for accuracy. For example, a trade-off for generalizability involves saying that the theory does not apply to the context of that result, and the theory will be accurate in contexts where it does apply. Alternatively, we sacrifice simplicity to restore accuracy, if we claim that the inconsistent result is due to a moderating contingency and once that moderator is considered, the result will be consistent with the theory.

The fact that such trade-offs are possible to 'save a theory from falsification' has been used to argue against falsification (Sørberg, 2005). Instead, CE views explicit discussion of such trade-offs as theory development. CE helps identify inaccuracies and make explicit what trade-offs are forced upon the theory. Theories (or research programs) with many such trade-offs are degenerate (Lakatos, 1970). CE prompts and documents such degeneration. CE is not about the next step of judging whether degeneracy is significant enough. Potential users of a theory can judge whether the theory lost too much simplicity or generalizability to be useful. For example, see Cross (1982) judging monetarism (a research program in macroeconomics) while explicitly reflecting on the Lakatosian ideas at the basis of that judgment.

Step 0 and step 1 contribute by identifying a lack of accuracy. Step 0 may seem difficult, but the same creativity that is displayed in thinking of alternative explanations should also allow us to reinterpret results from theories that oppose that result. Regarding step 1, developing a CE does not require fully fleshed-out theories. Instead, CEs consist of the most straightforward, and commonly held, associations of concepts (i.e. accepted propositions in Davis, 1971). The role of the following steps is to

specify which association to make less straightforward; this is the complication where the theory's simplicity or generalizability is sacrificed for accuracy. That process of complication leads to the theory becoming more fleshed out rather than that a fully fleshed-out theory is required in step 1. Still, CEs in step 1 must have a level of explicitness, detail, and connection to literature more similar to hypothesis development than to the generally weak argumentation for alternative explanations (Spector & Brannick, 2011).

Steps 2 to 6 contribute by identifying ways to restore accuracy by trading-off simplicity and/or generalizability. Thus, one of the accepted propositions is negated and replaced by a proposition that is more complex and more 'interesting' (Davis, 1971). There may be cases where you have an intuition that a theory implies the opposite of a result, so you have found a CE, but then upon further reflection the implication is not so straightforward. For example, a gravitational theory predicts the location of a planet, but a result shows that the planet is not at that location. While writing the CE you discover that your intuition was simplistic; the theory only predicted that location under the assumption that there was no other nearby planet pulling the focal planet away from its orbit. In that case, it is tempting to scratch the CE. However, CE values making explicit this step of further reflection; showing the reader where the intuition needs to be complicated. Thus, a CE author may decide that many readers would have the same intuition, so explaining that complication is a valuable contribution. As another example, step 5 prompts CE authors to also present the second and third best challenges they came up with. For instance, maybe you challenged a premise by adding the complication that that premise only applies to gaseous planets (and the result was found for a solid planet). CE encourages including that challenge, even if another challenge ends up being more plausible.

Step 7 makes summarizes the complication; making explicit the degeneration that is forced upon the theory by the result. It is possible that a result is inconsistent with a CE because of bad measures, auxiliary premises, or research designs. The CE author can decide whether a CE with a step 7 that reads 'Measure X does not capture concept A (in some context)' con-

tributes enough to be worth the effort. If a challenged premise is paper-specific, the contribution may be small. On the other hand, the contribution may be large enough if measure X is typical. For example, see Cook et al. (1979) discussing why insignificant results regarding the cognitive bias ‘sleeping effect’ are due to operationalizations not appreciating theoretical nuances. Cook et al. (1979)’s paper is like a CE paper, except that CE reframes the discussion from ‘in this paper we remind people of some important nuances in the theory, which empirical studies have failed to appreciate, which led to insignificant results’ to ‘insignificant results have forced us to appreciate the importance of some nuances, and in this paper, we make explicit the nuances we now believe to be important’.

Full circle

Given those details, we can see how CE counteracts underreporting by improving meta-analyses. Meta-analyses use concept labels as inclusion criteria. For example, Heugens and Lander (2009)’s meta-analysis on ‘mimetic pressure’s effect on isomorphism’ uses a variety of concept labels to search for literature (e.g., ‘isomorphism’, ‘institutional theory’). Underreporting practices cause studies to be described in terms of concepts that are supported by the result. Therefore, meta-analyses disproportionately include studies that support the theory (Murphy & Aguinis, 2019). (Note also that while HARKing (Hollenbeck & Wright, 2017) and abduction (Locke et al., 2008; Schwab & Starbuck, 2017) do not mislead like HARKing, they still lead studies to be described in terms of concepts that are supported by the results).

Enter CE. For example, a study finds a positive effect of ‘competition’ (measured as the number of firms in the same industry) on ‘differentiation’. A CE for that result is that a greater number of firms in the same industry can be interpreted as mimetic pressure (e.g., Haveeman, 1993) and differentiation is the opposite of isomorphism. Institutional theory suggests that mimetic pressure should increase isomorphism. Therefore, mimetic isomorphism theory implies a negative effect of the number of firms in an industry and differentiation; i.e., the opposite of the finding. Before the CE, this paper ‘about differentiation’ would fall outside of

Heugens and Lander’s meta-analysis inclusion criteria, so it would (systematically) fail to include results like these opposite to the theory. By contrast, after the CE is published, the result is described using theories that are not supported by it, so meta-analyses would include it. HARKing and not writing up tests could continue at the usual rate, but with CE, the proportion of rejected hypotheses among reported evidence would be greater (and closer to the true proportion).

Conclusion

Proposals to combat underreporting focus on preventing underreporting practices; e.g., de-emphasize p-values (Bettis, 2012), stop using “ $p < 0.05$ ”, (Bettis, 2012; Wasserstein et al., 2019), or abandoning null-hypothesis significance testing (Schwab et al., 2011). Such proposals are less feasible due to the inertia of current practice. By contrast, CE counteracts underreporting without preventing practices that lead to underreporting; it adds to, rather than changes, current practice. That increases feasibility. That is why the explanation of CE’s value can assume that HARKing and not writing up tests continue at the usual rate. CE’s value does not depend on whether a hypothesis was truly made after the results were known, nor does its value depend on what caused underreporting. CE makes use of the potential for interacting with studies after publication.

Moreover, CE is a useful practice, even if underreporting did not exist. First, CE also increases research reliability more directly. When CE is done by others than those who found the result, research reliability is increased simply by having an extra person thinking through the meaning of the data from a fresh perspective. Second, we put strain on others when collecting data. This comes with a responsibility to make the most of our data. CE helps fulfill that responsibility by reusing published results, in contrast to demands for efficient rather than comprehensive presentation, and novel findings.

In sum, I hope people use CE to learn more from the same findings and especially learn about, and from, those things that we currently miss due to underreporting.

Acknowledgements

I would to thank the reviewers, Pablo Martin de Holan, Jana Retkowsky, Arjen van Witteloostuijn, and the OT reading group at Tilburg University for their encouragement and valuable feedback on an earlier version of this work.

References

- Bedeian, A., Taylor, S., & Miller, A. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education*, 9(4), 715–725 (see p. 1).
- Bettis, R. A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal*, 33(1), 108–113. <https://doi.org/10.1002/smj.975> (see p. 5)
- Cook, T. D., Gruder, C. L., Hennigan, K. M., & Flay, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin*, 86(4), 662–679. <https://doi.org/10.1037/0033-2909.86.4.662> (see p. 5)
- Cross, R. (1982). The Duhem-Quine Thesis, Lakatos and the Appraisal of Theories in Macroeconomics. *The Economic Journal*, 92(366), 320. <https://doi.org/10.2307/2232443> (see pp. 2, 4)
- Davis, M. S. (1971). That's Interesting!: Towards a Phenomenology of Sociology and a Sociology of Phenomenology. *Philosophy of the Social Sciences*, 1(2), 309–344. <https://doi.org/10.1177/004839317100100211> (see pp. 2, 4)
- Haveman, H. A. (1993). Follow the Leader: Mimetic Isomorphism and Entry Into New Markets. *Administrative Science Quarterly*, 38(4), 593. <https://doi.org/10.2307/2393338> (see p. 5)
- Heugens, P. P. M. A. R., & Lander, M. W. (2009). Structure! Agency! (And Other Quarrels): A Meta-Analysis Of Institutional Theories Of Organization. *Academy of Management Journal*, 52(1), 61–85. <https://doi.org/10.5465/amj.2009.36461835> (see p. 5)
- Hines, R. (1988). Popper's methodology of falsificationism and accounting research. *The Accounting Review*, 63(4), 657–662 (see p. 4).
- Hollenbeck, J. R., & Wright, P. M. (2017). Harking, Sharking, and Tharking: Making the Case for Post Hoc Analysis of Scientific Data. *Journal of Management*, 43(1), 5–18. <https://doi.org/10.1177/0149206316679487> (see pp. 3, 5)
- Johns, G. (2019). GUIDEPOST: Departures from conventional wisdom: Where's the next opposite effect? *Academy of Management Discoveries*. <https://doi.org/10.5465/amd.2019.0226..> (See p. 1)
- Lakatos, I. (1970, September 2). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge* (1st ed., pp. 91–196). Cambridge University Press. <https://doi.org/10.1017/CBO9781139171434.009>. (See pp. 2, 4)
- Locke, K., Golden-Biddle, K., & Feldman, M. S. (2008). Perspective—Making Doubt Generative: Rethinking the Role of Doubt in the Research Process. *Organization Science*, 19(6), 907–918. <https://doi.org/10.1287/orsc.1080.0398> (see pp. 3, 5)
- Mazzola, J. J., & Deuling, J. K. (2013). Forgetting What We Learned as Graduate Students: HARKing and Selective Outcome Reporting in I-O Journal Articles. *Industrial and Organizational Psychology*, 6(3), 279–284. <https://doi.org/10.1111/iops.12049> (see p. 1)
- Murphy, K. R., & Aguinis, H. (2019). HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results? *Journal of Business and Psychology*, 34(1), 1–17. <https://doi.org/10.1007/s10869-017-9524-7> (see p. 5)
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, 349(6251) (see p. 1).
- Rubin, M. (2017). When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress. *Review of General Psychology*, 21(4), 308–320. <https://doi.org/10.1037/gpr0000128> (see pp. 1, 3)
- Schwab, A., & Starbuck, W. (2012). Using baseline models to improve theories about emerging markets. In C. Wang, D. Ketchen, & D. Bergh (Eds.), *West meets east: Toward methodological exchange (research methodology in strategy and management)* (pp. 3–33). Emerald Group Publishing Limited. (See p. 3).
- Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). PERSPECTIVE—Researchers Should Make Thoughtful Assessments Instead of Null-Hypothesis Significance Tests. *Organization Science*, 22(4), 1105–1120. <https://doi.org/10.1287/orsc.1100.0557> (see p. 5)
- Schwab, A., & Starbuck, W. H. (2017). A Call for Openness in Research Reporting: How to Turn Covert Practices Into Helpful Tools. *Academy of Management Learning & Education*, 16(1), 125–141. <https://doi.org/10.5465/amle.2016.0039> (see pp. 3, 5)
- Søberg, M. (2005). The Duhem Quine thesis and experimental economics: A reinterpretation. *Journal of Economic Methodology*, 12(4), 581–597. <https://doi.org/10.1080/13501780500343680> (see p. 4)
- Spector, P. E., & Brannick, M. T. (2011). Method-

- ological Urban Legends: The Misuse of Statistical Control Variables. *Organizational Research Methods*, 14(2), 287–305. <https://doi.org/10.1177/1094428110369842> (see p. 4)
- van Hugten, J., & van Witteloostuijn, A. (2021). The state of the art of hypothesis testing in the social sciences. In H. Mandele & A. van Witteloostuijn (Eds.), *A future for economics* (1st, pp. 167–185). VU University Press. (See p. 1).
- van Witteloostuijn, A. (2016). What happened to Popperian falsification? Publishing neutral and negative findings: Moving away from biased publication practices (A. Klarsfeld, L. C. Ng, & S. Eddy, Eds.). *Cross Cultural & Strategic Management*, 23(3), 481–508. <https://doi.org/10.1108/CCSM-03-2016-0084> (see p. 1)
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond “ $p < 0.05$ ”. *The American Statistician*, 73, 1–19. <https://doi.org/10.1080/00031305.2019.1583913> (see pp. 3, 5)
- Weick, K. (1999). Conclusion: Theory construction as disciplined reflexivity: Tradeoffs in the 90s. *The Academy of Management Review*, 24(4), 797–806 (see p. 4).