Dul, J., van der Laan, E., Kuik, R, & Karwowski, M. (2019). Necessary Condition Analysis: Type I error, power, and over-interpretation of test results. A reply to two comments on NCA. Working paper. A short version of this working paper is published in *Frontiers in Psychology*.

Necessary Condition Analysis: Type I error, power, and over-interpretation of test results. A reply to two comments on NCA.

Jan Dul[1*], Erwin van der Laan[1], Roelof Kuik[1], Maciej Karwowski[2]

[1]Rotterdam School of Management, Erasmus University, Rotterdam, the Netherlands

[2]Institute of Psychology, University of Wrocław, Wrocław, Poland

*** Correspondence:**
Corresponding Author
jdul@rsm.nl

Keywords: Necessary Condition Analysis, NCA, null hypothesis testing, alternative hypothesis, significance, power, type 1 error, *p*-value

Abstract

This note is a reply to two recent articles by Sorjonen and colleagues in *Frontiers in Psychology* that comment on Necessary Condition Analysis (NCA) and its statistical significance test. The first comment states that NCA does not protect the NCA researcher from making type I error. The second comment states that NCA's significance test is powerful but does not test necessity. We argue that NCA's test is powerful and protects the researcher against type I error (as statistically defined), and that - contrary to Sorjonen & Melin's suggestion - testing the research hypothesis, cannot be expected from any null hypothesis test, including not from NCA's test.

1 Introduction

Necessary Condition Analysis (NCA) is a novel data analysis approach that is based on necessity logic (Dul, 2016). We welcome the two comments on NCA from Sorjonen and colleagues (Sorjonen, Wikström & Melin, 2017; Sorjonen & Melin, 2019), which contribute to the scientific debate and to the advancement and proper use of NCA. NCA predicts the absence of an outcome when the condition is absent, rather than the presence of the outcome when the condition is present. To quantify this effect, NCA draws a ceiling line on top of the data in an XY scatter plot. The ceiling line represents the level of X that is necessary but not sufficient for reaching a given level of Y[i]. The empty space above the ceiling line is the necessity effect size. This empty space characterizes the constraint that X puts on Y. The lower the line, the larger the empty space and the higher the required X for a given Y. For example, in this journal, Shi, Wang, Yang, Zhang and Xu (2017) used NCA to show that intelligence is necessary for creativity, replicating previous findings (Karwowski, Dul, Gralewski, Jauk, Jankowska, Gajda, Chruszczewski, & Benedek, 2016).

2 Comment 1: randomness of the empty space and Type I error

The first comment on NCA (Dul, 2016) and one of its earliest applications in psychology (Karwowski et al., 2016) is the article 'Necessity as a function of skewness' by Sorjonen et al. (2017). Sorjonen et al. (2017) simulate that an empty space in the upper left corner can be produced

by unrelated variables with skewed distributions, hence not only by a necessity relationship between X and Y. This observation is correct. However, the randomness of the empty space can be detected by NCA's significance test (Dul, van der Laan, & Kuik, 2019). NCA's significance test is a null hypothesis test, assuming that the variables are unrelated. Specifically, NCA's test is a permutation test that produces a sampling distribution that represents the null hypothesis[ii]. The permutation test, and NCA's significance test, are also called a 'randomness tests'. Although both 'significance test' and 'randomness test' are used names for NCA's null hypothesis test, in this note we will call the test 'randomness test' because this explicitly reflects the purpose of the test. The test consist of the following steps (Dul et al., in press, p. 4): "Step 1: Calculate the necessity effect size for the sample. Step 2: Formulate the null hypothesis, which states that X and Y in the population are not related. Step 3: Create a large random set of resamples (e.g., 10,000) using approximate permutation. Step 4: Calculate the effect size of all resamples. Step 5: Compare the estimated distribution of effect sizes of random resamples with the effect size of the observed sample (see Step 1). The fraction of random resamples for which the effect size is equal to or greater than the observed effect size (*p*-value) informs us about the statistical (in)compatibility of the data with the null hypothesis."

Dul et al. (2019) show by simulations and by referring to mathematical proofs that NCA's randomness test can identify whether an empty space may be due to random chance or not. The test "…is intended to answer the question: 'Can the observed effect size be the result of random chance?' by responding: 'Yes, but with probability smaller than *p*.'" (Dul et al., 2019, p.2). When researchers use the test for making a binary decision about the null hypothesis, a low actual *p*-value (e.g. smaller that the threshold *p*-value of 0.05) indicates that that null hypothesis should be rejected, and a high actual *p*-value (e.g. greater than the threshold *p*-value of 0.05) indicates that that null hypothesis should *not* be rejected. The latter outcome helps researchers to avoid making type I error: rejecting the null hypothesis when the null is true.

3 Comment 2: Type I error, power, and accepting an alternative

The second comment on NCA is the article 'Predicting the significance of necessity' by Sorjonen & Melin (2019). In this study, three simulations are performed. In the first simulation the population has a true necessity effect (upper left corner is empty), in the second simulation the population has true necessity effect and a true sufficiency effect (the upper left corner is empty and the lower right corner is empty), and in the third simulation the population has a true sufficiency effect (lower right corner is empty). The simulations are performed with different values of the necessity and/or sufficiency effect sizes (different sizes of the corresponding empty spaces). For each simulation the performance of NCA's randomness test is evaluated as follows: "The objective of the present study was to evaluate if and how calculated *p*-values for necessity effects can be predicted from true population necessity effect (i.e., amount of empty space in the upper-left corner in an X–Y-plot), true population sufficiency effect, and sample size. This gives indications of the power of the method as well as risk for type 1-errors." (Sorjonen & Melin, 2019, p. 2). Type I error is defined as the chance of rejecting the null hypothesis when the null is true. Power is defined as the chance of rejecting the null hypothesis when the alternative is true. Although the validity of the NCA's randomness test regarding type I error has already been proven (see above), Sorjonen & Melin (2019) re-evaluate this quality. In addition, they evaluate the power of the test for the first time. Furthermore, they discuss their simulation results in terms of ability of NCA's test to test the alternative necessity hypothesis. Below we discuss Sorjonen & Melin's (2019) interpretations of the findings.

3.1 Type I error

Sorjonen & Melin (2019) calculate actual *p*-values (with NCA's randomness test) for the three simulated relationships between X and Y outside the null hypothesis: necessity relation (H1: alternative 1), necessity and sufficiency relation (H2: alternative 2), and sufficiency relation (H3: alternative 3). Given the definition of type I error (rejecting a true null), type I error can only be evaluated when the null is true (H0 = null hypothesis). The null being true (randomness, un-relatedness of X and Y) implies the non-emptiness of all corners in the XY scatter plot. This is the situation that occurs in the three simulations when the necessity and sufficiency effect sizes are zero. Inspection of Sorjonen & Melin's (2019) simulation results indeed shows that when all effect sizes are zero and thus X and Y are unrelated (null is true), NCA's test correctly identifies randomness. With a chosen threshold value of $p = 0.05$, NCA's randomness test shows that the null, as expected, is not rejected in approximately 95% of the samples. The corresponding type I error rate is 5%, thus in approximately 5% of the samples the true null is rejected. Sorjonen & Melin (2019, p.5) seem to acknowledge this quality of NCA's randomness test: "Without any true population sufficiency effect, NCA did not seem to result in more type 1-errors than expected, i.e., 5%", at least for the case that the necessity effect is also absent (ensuring that the null is true). Hence, the conclusion remains that when X and Y are unrelated, NCA's randomness test is valid.

3.2 Power

Sorjonen & Melin (2019) also have results on the power of the test for several simulated relationships between X and Y. Given the definition of power (rejecting the null when the alternative is true), the correctness of the prediction of power can only be evaluated when the alternative is true. The alternative of the null is that X and Y are *not* unrelated, thus are related. In Sorjonen & Melin's (2019) simulation three specifications of the alternative are evaluated: necessity (H1), sufficiency (H3), and both (H2). When an alternative is true, the actual *p*-value should be small. The simulation results indeed show that the actual *p*-value rapidly approaches zero when necessity and/or sufficiency effect sizes increase. Hence, NCA's randomness test is not only valid regarding type I error but has also high power. Sorjonen & Melin (2019, p. 3) seem to acknowledge also this quality of NCA's randomness test: "this apparent high power of NCA could be seen as a positive characteristic."

3.3 Over-interpretation of the null hypothesis test result

The value of Sorjonen & Melin's article is not only that they show that NCA's randomness test can handle type I error and has power. They also show that over-interpretation of the test results of a null hypothesis test like NCA's randomness test can be dangerous. Sorjonen & Melin (2019) over-interpret the test results, namely that a rejection of the null hypothesis can be understood as the acceptance of a *specific* alternative hypothesis, in this case necessity. When introducing NCA's test, Dul et al. (2019, p. 1) had the following goal: "We propose a statistical significance test that evaluates the evidence against the null hypothesis of an effect being due to chance. Such a randomness test helps protect researchers from making Type 1 errors and drawing false positive conclusions". We warned against over-expectation regarding the test: "just like any other statistical method that uses the *p*-value for statistical inference, the proposed approximate permutation test has all the limitations of the *p*-value". One of these limitations and its related over-interpretation is the one discussed in Sorjonen & Melin's article and that is formulated by Szucs & Ioannidis (2017, p 8.) as follows: "A widespread misconception … is that rejecting H0 allows for accepting a *specific* H1 .... This is what most practicing researchers do in practice when they reject H0 and argue for their specific H1 in turn" [emphasis in the original]. When introducing NCA's randomness test we might have been more explicit about this limitation of any null hypothesis test, including NCA's, and using two names for the same test ("significance test" and randomness test") may have confused some

readers. So let us be clear: NCA's test is a test of the null hypothesis, not a test of the alternative hypothesis. The test does not attempt to prove necessity, but instead intends to reject or not reject randomness of the empty space. Not rejecting randomness is *not* the same as accepting necessity. The correct interpretation of a test result of actual $p \leq 0.05$ is that the result is not due to random chance, no more, no less.

Following the described common mis-conception of the null and its test, Sorjonen & Melin's (2019) article focusses on NCA's obvious inability to prove necessity when the null is rejected. In Sorjonen & Melin's discussion of the power of the test (the probability of rejecting the null when the alternative is true), they only use necessity as the alternative hypothesis (H1). But the test also rejects, and should reject, the null when the sufficiency alternative is true (H2 and H3). Sorjonen & Melin (2019) do not mention this as another indication of the power of the test. Instead they call this latter result a 'type 1 error'. For example, they state (p. 3) that "While sample size had no effect on the probability to get a significant observed necessity effect, i.e., the risk for type 1-error, this risk increased with increased true population sufficiency effect." Note that this interpretation of "type I error" does not correspond to the definition in statistics (the probability of rejecting a true null), which is only defined when the *null* is true, not when an alternative is true. Also in the discussion section Sorjonen & Melin (2019) focus on the incorrect over-interpretation of the test results. When referring to the high power of NCA's test, they state: "However, one might also become a bit worried by the ease with which people wanting to claim that X is a necessary condition for Y can overcome the obstacle of significance". In this worry, they assume that people make the incorrect over-interpretation of having a significant (actual $p \leq 0.05$) *necessity* result, whereas they truly have found a significant (actual $p \leq 0.05$) *non-random* result.

Sorjonen & Melin's (2019) focus on the common misinterpretation of the null hypothesis is unfortunate because it obscures their contribution: showing that NCA's randomness test is not only a valid test to identify that an empty space may be due to random chance, but is also a valid test regarding its ability to identify that an empty space is *not* due to random chance.

4 Conclusion

NCA's randomness test is a valid and powerful test to test the randomness of an empty space in the upper left corner of a XY scatter plot. A high actual *p*-value suggests that the result may be due to random chance and a low actual *p*-value suggests that the result may not be due to random chance. However, such low actual *p*-value cannot be interpreted as a proof of necessity. Rather, a low actual *p*-value can be considered as a "minimum statistical test" (Dul et al. 2019, p. 8), thus as a necessary but not sufficient condition for interpreting the empty space as being due to necessity. Sorjonen & Melin's article shows again that results of null hypothesis tests should not be over-interpreted.

It may seem disappointing that a null hypothesis test like NCA's randomness test can only test whether a result is due to randomness or not, and cannot test for a specific alternative hypothesis. However, this is inherent to null hypothesis testing. For direct testing of a necessity hypothesis, other statistical approaches need to be developed, such as Bayesian approaches. Such approaches are currently not available for NCA and may be a topic for future research.

5 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6 Author Contributions

JD wrote the first draft and revisions of the manuscript. RK, EvdL and MK contributed to successive revisions.

7 Acknowledgments

We thank Benjamin Krebs and Henk van Rhee for their suggestions.

8 References

Dul, J. (2016). Necessary Condition Analysis (NCA): logic and methodology of "necessary but not sufficient" causality. *Organizational Research Methods* 19, 10–52.

Dul, J., van der Laan, E., & Kuik, R. (2019). A statistical significance test for Necessary Condition Analysis. *Organizational Research Methods* (in press).

Karwowski, M., Dul, J., Gralewski, J., Jauk, E., Jankowska, D. M., Gajda, A., Chruszczewski, M. H. & Benedek, M. (2016). Is creativity without intelligence possible? A necessary condition analysis. *Intelligence, 57*, 105-117.

Shi, B., Wang, L., Yang, J., Zhang, M., & Xu, L. (2017). Relationship between Divergent Thinking and Intelligence: An Empirical Study of the Threshold Hypothesis with Chinese Children. *Frontiers in Psychology*, 8, 254.

Sorjonen, K., & Melin, B. (2019). Predicting the significance of necessity. *Frontiers in psychology*, 10, 283.

Sorjonen, K., Wikström Alex, J., & Melin, B. (2017). Necessity as a Function of Skewness. *Frontiers in psychology*, 8, 2192.

Szucs, D., & Ioannidis, J. (2017). When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in human neuroscience*, 11, 390.

---

[i] The ceiling line has observations below it but not above it. The ceiling line represents therefore that a certain X value is necessary but not sufficient for a certain Y value on the ceiling line. It is possible that there is also a floor line. The floor line has observations above it but not below it. The floor line represents that a certain X value is sufficient but not necessary for a certain Y value on the floor line. It is not 'paradoxical, as suggested by Sorjonen & Melin (2019), that a ceiling line and a floor line are both present at the same time: the ceiling line still represents that a certain X value is necessary but not sufficient for a certain Y value on the ceiling line, also if that X were sufficient but not necessary for a (lower) Y value on the floor line.

[ii] The null sampling distribution is obtained by shuffling Y values over X values, or by shuffling X values Y values, which, contrary to what Sorjonen & Melin (2019) claim, gives identical results.