Dul, J., van der Laan, E., Kuik, R. (2018) A statistical significance test for Necessary Condition Analysis. Organizational Research Methods (in press)

A Statistical Significance Test for Necessary Condition Analysis

Jan Dul, Erwin van der Laan, Roelof Kuik

Rotterdam School of Management, Erasmus University, the Netherlands

Burgemeester Oudlaan 50

3062 PA Rotterdam, The Netherlands

Tel: +31 (0)10 408 1719

Correspondence to jdul@rsm.nl

Acknowledgments:

We are grateful to Benjamin Krebs, Fabian Nullmeier, and Henk van Rhee for providing valuable comments on earlier versions of this paper, and to the authors of Van der Valk et al. (2016) for providing their data set for re-analysis.

Abstract

In this article, we present a statistical significance test for necessary conditions. This is an elaboration of the Necessary Condition Analysis (NCA), which is a data analysis approach that estimates the necessity effect size of a condition X for an outcome Y. NCA puts a ceiling on the data, representing the level of X that is necessary (but not sufficient) for a given level of Y. The empty space above the ceiling relative to the total empirical space characterizes the necessity effect size. 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 to protect researchers from making Type 1 errors and drawing false positive conclusions. The test is an 'approximate permutation test'. The test is available in NCA software for R. We provide suggestions for further statistical development of NCA.

.

Keywords: Null hypothesis testing, permutation test, Necessary Condition Analysis, statistical significance, p-value

A Statistical Significance Test for Necessary Condition Analysis

Necessary Condition Analysis (Dul, 2016) is as a tool for researchers to develop and test necessary, but not sufficient conditions. A necessary condition enables the outcome when present and constrains the outcome when absent. NCA assumes that outcome Y is bound by condition X by drawing a ceiling line on top of the data in an XY scatter plot. The line defines the empty space in the upper left corner of the scatter plot[i]. This empty space suggests that high values of Y are not possible with low values of X and indicates that X constrains Y. The size of the empty space relative to the total space with observations reflects the *extent* of the constraint that X poses on Y: the larger the empty space, the more X constrains Y. The necessity effect size (*d*) is the size of the empty space above the ceiling as a fraction of the total space where cases are observed or could be observed given by the minimum and maximum empirical or theoretical values of X and Y ('scope'[ii]). NCA's effect size *d* has values between 0 and 1.

The NCA effect size has been used in various organizational studies for testing the necessary condition hypothesis that 'X is necessary for Y'. For example, Van der Valk, Sumo, Dul, and Schroeder (2016) test whether trust and contracts are necessary for successful collaboration between buyers and suppliers for innovation. Arenius, Engel, and Klyver (2017) test whether particular gestation activities for establishing a new firm are necessary for profit two years after the firm's start, and Karwowski et al. (2016), Karwowski, Kaufman, Lebuda, Szumski, and Firkowska-Mankiewicz (2017) and Shi, Wang, Yang, Zhang, and Xu (2017) test the hypothesis that intelligence is necessary for creativity. Currently, such necessity hypotheses are assessed based on NCA effect size. A 0.1 threshold level for effect size is often applied,

hence a hypothesis is considered to be supported if the empty space above the data is at least 10% of the scope. However, testing a necessary condition hypothesis only based on effect size may produce unjustified conclusions, as the result may not be *statistically* significant. The observed necessity empty space may be caused by random chance. For example, when low X-values and high Y-values are relatively rare, an empty space is likely but may not be the result of necessity. This can happen when X and Y are unrelated random variables with normal or skewed distributions, which is not uncommon in the organizational sciences. Therefore, there is a need to protect the researcher who applies NCA against a Type 1 error: concluding that the empty space represents necessity, when it is actually a random occurrence.

In this article, we presume that the reader is familiar with NCA (Dul 2016). We advance the NCA's hypothesis testing approach by proposing a statistical significance test for testing the randomness of the effect size. Specifically, we provide a permutation test for NCA users to calculate the p-value. 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.' We demonstrate the application of the test with an example dataset and use Monte Carlo simulations to show that the permutation approach is a generic and valid randomness test. We provide suggestions for further statistical development of NCA.

Permutation test

Since Fisher (1935), statisticians have used the permutation test for statistical significance testing. Until recently, the test was not popular due to high computational demands (Hayes, 1996; Ludbrook & Dudley, 1998). Since the availability of fast computers, permutation tests have been developed for correlation and regression (Anderson & Robinson, 2001; DiCiccio &

Romano, 2017), ANOVA (Anderson, 2001), the General Linear Model (Winkler, Ridgway, Webster, Smith, & Nichols, 2014), and Qualitative Comparative Analysis (Braumoeller, 2015). The permutation test produces a p-value. The test is particularly useful when analytical approaches to estimate the p-value are not available, or when assumptions for these approaches do not hold.

The p-value is the theoretical probability that the value of a test statistic that summarizes the observed sample data, e.g., the observed effect size, is equal to, or larger than the value of this test statistic when the null hypothesis is true. Significance tests including the permutation test usually employ a *reductio ad absurdum* argumentation. This means that the null hypothesis is formulated, which states that the data of the observed sample are the result of a random data generation process in the population where X and Y are unrelated: the null hypothesis. Next, the probability (p) that the effect size of the *observed* sample is equal to or larger than the effect size of random samples is calculated. If this probability is small (e.g., $p < 0.05$), it is concluded that the observed sample is unlikely the result of a random process of unrelated variables (the null hypothesis is rejected), suggesting support for an alternative hypothesis.

In the permutation test, a distribution of random samples is produced under the null hypothesis by reshuffling observed X and Y values of cases of the observed sample. This ensures that under the null hypothesis X and Y are not related and a possible effect size is due to random chance. Notice that resampling *X and Y values* by permutation aims to mimic the *null-hypothesis distribution*. This is different from resampling *cases* by bootstrapping which aims to mimic the *population distribution*. Applying standard bootstrapping to NCA would result in an invalid significance test where the null hypothesis is overly rejected[iii]. Specifically, permutation resamples are constructed by assigning observed Y-values to observed X-values to obtain all

possible combinations (permutations) of observed X and Y values. One way of achieving this in a bivariate dataset is by letting observed X-values have a fixed order, and then permutating Y-values. For the permutation test, no assumptions about the distribution of the data are required. All that is needed to make the test valid is that the distribution of the Y (that is permutated) is 'exchangeable' (for a discussion on exchangeability see Good, 2005). In an experimental study, we can assume exchangeability when cases are randomly allocated to groups. In an observational study, we can assume exchangeability when the sample is a random sample from the population, which are common assumptions for statistical inference. The permutation test is a valid test (Hoeffding, 1952; Kennedy, 1995). Lehman and Romano (1998, p. 633) provide formal proof of this in Theorem 15.2.

Table 1 illustrates how the permutation test resamples from the observed sample. Suppose the sample consists of only three cases, thus with three values for X and three values for Y. The first observed Y value (y_1) has three possibilities to be assigned to an X value (x_1, x_2, x_3), for the second observed Y value (y_2) two possibilities are left, and for the third observed Y value (y_3) one possibility is left, which results in six (3x2x1 = 3 factorial = 3!) possible resamples (permutations). One from all possible six permutations corresponds to the observed sample. When the observed effect size is the smallest of the six random samples, the proportion of random samples that has an effect size that is equal to or larger than the observed effect size equals 6/6 = 1 (p=1). When the observed effect size is the largest of the six random samples, the proportion of random samples that has an effect size that is equal to or larger than the observed effect size equals 1/6 = 0.17 (p=0.17)[iv].

INSERT TABLE 1 ABOUT HERE

The number of permutations rapidly increases with sample size. For example, a bivariate sample of 10 cases (n=10) results in 10! = 3,628,800 permutations, and a sample of 50 cases (which is considered a small sample in many organizational fields) results in 50! permutations, which is around 30,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000 permutations.

To handle this computational problem[v], a large subset of all possible permutations is randomly selected to approximate the permutation distribution. Such a permutation test is called an 'approximate permutation test'[vi], which produces an estimate of the exact p-value. An observed Y-value is randomly selected from all observed Y-values without replacement (because a permutation does not allow the same value of Y to be assigned to X-values more than once) and assigned to an observed X-value. This process is repeated until all observed X-values have a Y-value, hence the size of the resample equals the observed sample size. This procedure is repeated to obtain a large random set of constructed resamples. The test statistic is computed for each resample, and the distribution of random test statistic is compared to the observed sample value of the test statistic. The proportion of random resamples for which the value of the test statistic is equal to or larger than the observed value of the test statistic is the estimated p-value, which informs us about the statistical (in)compatibility of the data with the null hypothesis. We propose the following five steps for performing an approximate permutation test for NCA:

1. Calculate the necessity effect size for the sample. Presume that the sample is a random sample from the population for an observational study and that cases are

randomly allocated to groups for an experimental study. These are common assumption for statistical inference.

2. Formulate the null hypothesis, which states that X and Y in the population are not related: Prob(Y|X) ≡ Prob(Y). Under the null hypothesis, the theoretical mean necessity effect size $d = 0$, whereas the effect size under the null hypothesis is likely $d > 0$ for a specific finite sample. This effect size is a random effect arising from (finite) sampling.
3. Create a large random set of resamples (e.g., 10,000, see below) using approximate permutation.
4. Calculate the effect size of all resamples. The set of effect sizes comprises an estimated distribution of effect size under the null hypothesis that X and Y are not related.
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.

Demonstration

Example

For illustration, we applied the approximate permutation test to a dataset for testing the hypothesis that trust between companies is necessary for collaborative innovation performance. Van der Valk et al. (2016) studied buyer-supplier relations and used the NCA effect size to test

the hypothesis that trust is necessary for supplier-led innovation in collaborations between buyer and supplier firms. They studied 48 buyer-supplier service outsourcing collaborations. From the trust dimensions that were studied by Van der Valk et al. (2016) in the present paper we considered only goodwill trust. This trust dimension relates to the intention to fulfill an agreed role in the collaboration.

The approximate permutation test is applied to this example as follows:

1. Calculate the observed necessity effect size.

   If NCA's CE-FDH ceiling line[vii] is selected, the necessity effect size for trust is 0.31.

2. Formulate the null hypothesis.

   The null hypothesis states that trust and performance are not related and that any observed empty space in the upper left corner of the trust-performance scatterplot is due to random chance.

3. Create a large random set of permutation resamples.

   A performance value (Y) is randomly selected without replacement from the observed performance values and assigned to an observed trust value (X). This process is repeated 48 times (corresponding to the sample size) until all X-values have a Y-value, and a random resample is obtained. This procedure is repeated 10,000 times to obtain 10,000 random resamples.

4. Calculate the effect size of each resample.

   The CE-FDH effect size is calculated for all resamples.

5. Compare the distribution of effect sizes of the random resample with the observed effect size (Figure 1).

Figure 1 shows the distribution of the CE-FDH effect sizes of the 10,000 random samples.

INSERT FIGURE 1 ABOUT HERE

The observed effect size of 0.31 is larger than all but 17 random effect sizes. Hence, the probability that the random effect size is equal to or larger than the observed effect size is less than 17/10,000 (p<0.0017). The example shows that the observed effect size under the assumption that the null hypothesis is true is very rare, which is an indication that the null hypothesis (the observed effect size is due to random chance) does not explain the data, hence that the alternative hypothesis may be supported (trust is necessary for performance).

Monte Carlo simulation

We performed a Monte Carlo simulation to evaluate if the NCA approximate permutation test can correctly recognize an empty space as random chance, if the data generation process was random. Specifically, we built on the simulation study by Sorjonen, Akex and Melin (2017) who produced empty spaces in the upper left corner when X and Y were unrelated random variables with beta distributions. They repeatedly drew random samples from beta distributions with different values of skewness of X (X-skew), skewness of Y (Y-skew) and different sample sizes. The null hypothesis applied in all samples because X and Y were not related, and any effect size would be due to random chance.

INSERT FIGURE 2 ABOUT HERE

Figure 2 (left) shows results of the original simulation of Sorjonen et al. (2017) with the effect size on the vertical axes (using the CR-FDH ceiling line) and different values of X-skew on the horizontal axes. The nine plots have different values of Y-skew and sample size. Figure 2 (left) shows 7,350 dots and each dot is a sample. The plots show that the effect size can be large, up to more than 0.6. The effect size is larger for smaller sample size, more negative skewness of X (low values of X are rarer) and more positive skewness of Y (high values of Y are rarer).

We added the approximate permutation significance test to this simulation to verify if the test could detect that the effect sizes were due to random chance. Because of high computational demands when combining the permutation approach with the original simulation, we selected a relatively small number of permutations (500). Yet, the computation time for this simulation was about 15 hours. We calculated the p-value of the effect size for each sample. Figure 2 (right) presents the results of the estimation of the p-value using the NCA approximate permutation test. The plots on the right are the same as the plots on the left, except for the vertical axis: on the left, the vertical axis is the effect size and on the right, it is the estimated p-value for the effect size. The total number of dots (samples) is 7,350. The horizontal line in the plots on the right corresponds to a chosen threshold p-value of 0.05, thus considering sample effect sizes as random chance if $p > 0.05$. The results show 6,991 samples (out of 7,350) with NCA $p > 0.05$ (suggesting randomness). This is 95.1% of all samples. This illustrates that the NCA significance test can identify randomness under different conditions and with a probability corresponding to the threshold significance level pre-selected by the researcher (in this case 0.05). We found similar results with additional simulations with other distribution functions (uniform, truncated normal, triangle, not reported here). In other words, our simulation findings for the NCA effect

size are consistent with the general analytical demonstrations of the validity of the permutation test in the literature (e.g., Lehman and Romano, 1998).

Accuracy of the p-value estimation

If all possible permutation resamples were part of the sampling distribution, the p-value would be exact. In the approximate permutation test the p-value has some uncertainty. The exact p-value equals the estimated p-value plus or minus the accuracy of the estimated p-value (p-accuracy). The p-accuracy can be estimated because the estimated p-value is a proportion that has a binomial distribution. The p-accuracy depends on the number of permutations and on the estimated p-value. Inversely, one can determine the required number of permutations for a desired p-accuracy, which depends on the estimated p-value (Table 2)[viii].

INSERT TABLE 2 ABOUT HERE

Table 2 shows that with a large number of permutations the accuracy of the approximate permutation test is acceptable for the practical purposes of significance testing. The accuracy increases when increasing the number of permutations (within the limits of computation time). If the estimated p-value is 0.05, one can be confident (confidence level 95%) that the p-accuracy is around 0.004 with 10,000 permutations, with an exact p-value within the range of 0.046 to 0.054. P-accuracy is around 0.001 with 100,000 permutations, with an exact p-value within the range of 0.049 to 0.051. Hesterberg (2014, p. 81) recommends (somewhat arbitrarily) "… 10,000 permutations for routine use, and more when accuracy matters."

Discussion

The proposed statistical significance test for NCA is a relevant addition to the NCA effect size. The observed effect size may be the result of random chance. Hence, testing a hypothesis only based on an effect size threshold may be insufficient. Based only on effect size, the researcher may consider the alternative hypothesis as plausible, whilst the data fits the null hypothesis (false positive, due to Type 1 error). With the proposed approximate permutation test the NCA researcher has a tool to assess the randomness of the observed effect size. Observed effect sizes with p-values above 0.05 cast doubts about the statistical significance of the result.

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 (Wasserstein and Lazar, 2016). The p-value only provides indirect information about the evidence that an observed effect may be the result of random chance, and at best provides indirect support for the hypothesis of interest. Statistical testing of empirical data is a complex endeavor (Forstmeier, Wagenmakers, & Parker, 2016), and no universal method exists for statistical inference (Gigerenzer & Marewski, 2015). In the proposed p-test for NCA, we stay close to the original p-value approach as suggested by Fisher (1925). A small p-value is either a rare result that happens only with probability p (or lower) or is an indication that the null hypothesis does not explain the data.

Although the proposed statistical test for NCA is an important step forward, further statistical developments are needed. In the current ‘descriptive statistics’ phase of NCA development, the ceiling line and effect size just describe the data, and inferential statistics is limited to point estimates of NCA parameters. With the proposed approximate permutation test, we have entered the next phase of development: ‘statistical significance testing’, namely testing

effect sizes against a null hypothesis to avoid Type 1 error and false positives. In this phase, other null hypothesis testing approaches may be developed, such as testing approaches with assumptions about a relationship between X and Y under the null hypothesis, parametric analytical approaches based on assumptions of distributions of X and Y[ix], or bootstrapping approaches beyond standard bootstrapping. Analytical and bootstrapping approaches may be particularly useful for the next phase of development: 'standard error/confidence interval' estimations to provide interval estimates, namely developing a precision measure of the point estimates (which could also be used for significance testing). Further phases of statistical development of NCA could include more advanced approaches such as Bayesian approaches for directly testing the hypothesis of interest, instrumental variable approaches for checking assumed causal directions, and approaches that include modeling measurement error.

Researchers wishing to perform NCA are recommended to test the randomness of the observed effect size by using the approximate permutation test. This test can be considered as a minimum statistical test for NCA. At least three necessary but not sufficient conditions exist for a condition being a necessary condition: (1) theoretical justification, (2) effect size $d > 0$, and (3) small p-value (e.g., $p < 0.05$). The approximate permutation test is implemented in the NCA software for R, version 3.0 onwards (Dul, 2015).

References

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences*, 58(3), 626-639.

Anderson, M. J., & Robinson, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1), 75-88.

Arenius, P., Engel, Y., & Klyver, K. (2017). No particular action needed? A necessary condition analysis of gestation activities and firm emergence. *Journal of Business Venturing Insights*, 8, 87-92.

Braumoeller, B. F. (2015). Guarding against false positives in qualitative comparative analysis. *Political Analysis*, 23(4), 471-487.

DiCiccio, C.J., & Romano. J. P. (2017) Robust permutation tests for correlation and regression coefficients, *Journal of the American Statistical Association*, 112 (519), 1211-1220.

Dul, J. (2015). Necessary Condition Analysis (NCA) for R: A quick start guide. Retrieved from http://ssrn.com/abstract=2624981 or http://repub.eur.nl/pub/78323/.

Dul, J. (2016) Necessary Condition Analysis (NCA): Logic and methodology of "necessary but not sufficient" causality. *Organizational Research Methods,* 19(1), 10-52.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2016). Detecting and avoiding likely false□positive findings–a practical guide. *Biological Reviews*.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, 41(2), 421-440.

Good, P. (2005). *Permutation tests: a practical guide to resampling methods for testing hypotheses.* Springer Science & Business Media.

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing H_0: ρ= 0. *Psychological Methods*, 1(2), 184.

Hesterberg, T.C. (2014). What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum, available at http://arxiv.org/abs/1411.5279

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations, *The Annals of Mathematical Statistics*, 23, 169–192.

Karwowski, M., Dul, J., Gralewski, J., Jauk, E., Jankowska, D. M., Gajda, A…, … Benedek, M. (2016). Is creativity without intelligence possible? A Necessary Condition Analysis. *Intelligence*, 57, 105–117.

Karwowski, M., Kaufman, J. C., Lebuda, I., Szumski, G., & Firkowska-Mankiewicz, A. (2017). Intelligence in childhood and creative achievements in middle-age: The necessary condition approach. *Intelligence*, 64, 36-44.

Kennedy, P.E. (1995). Randomization Tests in Econometrics. *Journal of Business & Economic Statistics,* 13 (1), 85-94.

Lehman, E.L., & Romano, J.P. (1998). *Testing Statistical Hypotheses*. Springer.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*, 52(2), 127-132.

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., Alex, J. W., & Melin, B. (2017). Necessity as a Function of Skewness. *Frontiers in Psychology*, 8: 2192.

Valk, W. van der, Sumo, R., Dul, J. & Schroeder, R. (2016). When are contracts and trust necessary for innovation in buyer-supplier relationships? A Necessary Condition Analysis. *Journal of Purchasing and Supply Management*, 22(4), 266-277.

Wasserstein, R.L., & Lazar, N.A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70:2, 129-133.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage*, 92, 381-397.

Tables

Table 1

The six possible permutations (samples) for three cases (A,B,C) and two variables (X,Y)

Sample 1			Sample 2			Sample 3			Sample 4			Sample 5			Sample 6		
Case	X	Y	Case	X	Y	Case	X	Y	Case	X	Y	Case	X	Y	Case	X	Y
A	x_1	y_1	A	x_1	y_2	A	x_1	y_3	A	x_1	y_1	A	x_1	y_2	A	x_1	y_3
B	x_2	y_2	B	x_2	y_3	B	x_2	y_1	B	x_2	y_3	B	x_2	y_1	B	x_2	y_2
C	x_3	y_3	C	x_3	y_1	C	x_3	y_2	C	x_3	y_2	C	x_3	y_3	C	x_3	y_1

Table 2

Accuracy of p-value (p-accuracy) estimated by approximate permutation as a function of number of permutations (bold rows names) and estimated p-value (bold headers) (95% confidence that exact p-value = estimated p-value ± p-accuracy). For example, if the estimated p-value is 0.05, the exact p-value for 10,000 permutations lies with 95% confidence within the range from 0.046 to 0.054, and for 100,000 permutations lies within the range from 0.049 to 0.051.

Permutations	**p=0.2**	**p=0.1**	**p=0.05**	**p=0.01**	**p=0.005**	**p=0.001**	**p=0.0005**	**p=0.0001**
500	0.035	0.026	0.019	0.009	0.006	0.003	0.0020	0.0009
1,000	0.025	0.019	0.014	0.006	0.004	0.002	0.0014	0.0006
5,000	0.011	0.008	0.006	0.003	0.002	0.001	0.0006	0.0003
10,000	0.008	0.006	0.004	0.002	0.001	0.001	0.0004	0.0002
50,000	0.004	0.002	0.002	0.001	0.001	0.000	0.0002	0.0001
100,000	0.002	0.002	0.001	0.001	0.000	0.000	0.0001	0.0001
500,000	0.001	0.001	0.001	0.000	0.000	0.000	0.0000	0.0000
1,000,000	0.001	0.001	0.000	0.000	0.000	0.000	0.0000	0.0000

Figures

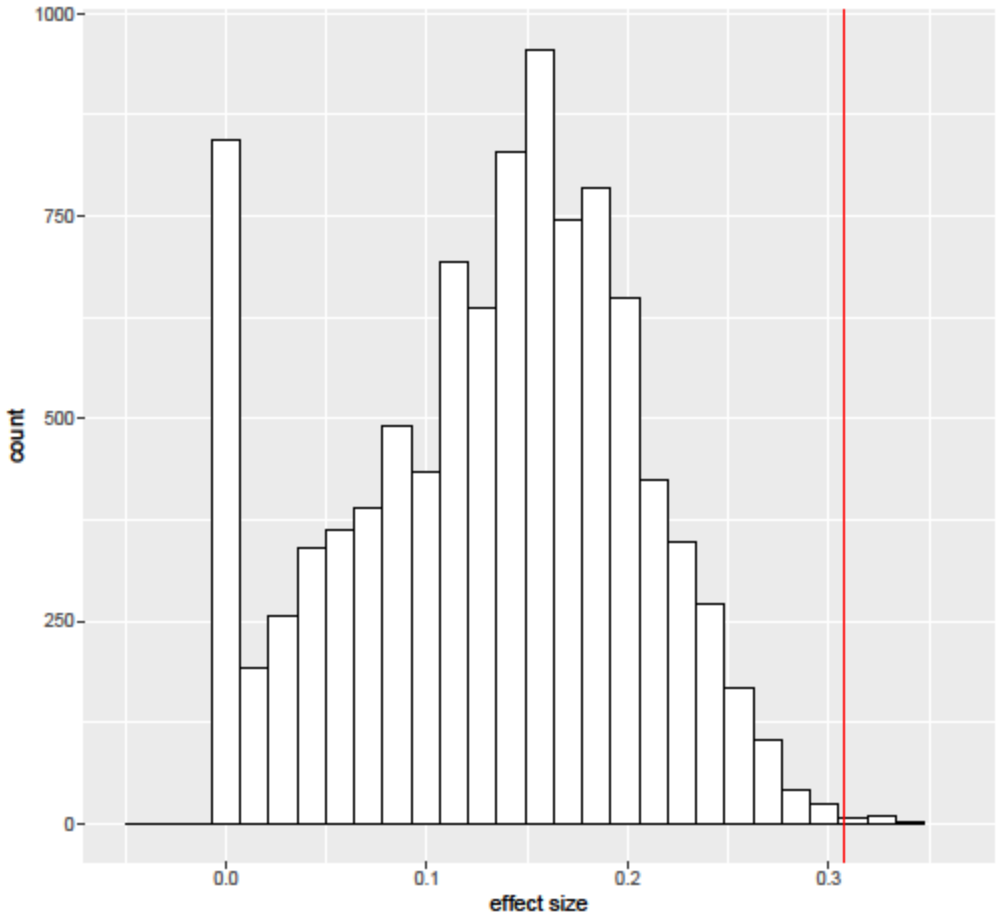

Figure 1. Statistical significance test with the null hypothesis stating that trust is not necessary for performance. (Data from Van der Valk et al. 2016). Distribution of necessity effect sizes (calculated with the CE-FDH ceiling line) under the null-hypothesis for 10,000 random samples generated by approximate permutation. Horizontal axis: effect size. Vertical axis: number of samples. Seventeen of the 10,000 random effect sizes are equal to or greater than the observed effect size (d = 0.31, p=0.0017), suggesting that the data do not fit the null hypothesis well.

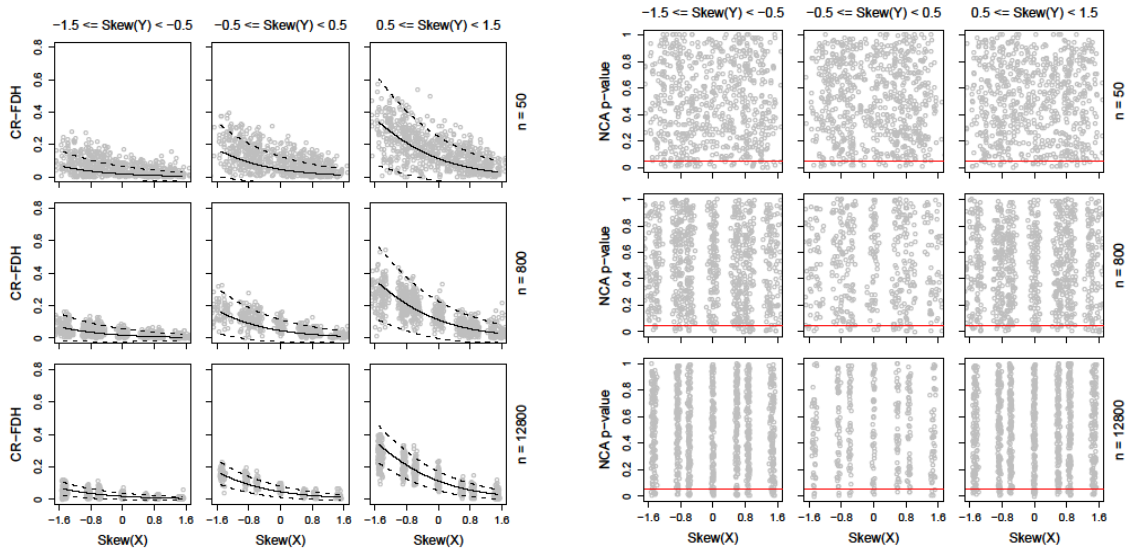

Figure 2. Monte Carlo simulation of CR-FDH effect size under the null hypothesis of no relation between X and Y, and effect size due to random chance (left), and corresponding p-values (right). Each dot is a sample. In both plots the total number of samples is 7,350.

Left: Simulation with different values of sample size (n), X-skew, and Y-skew (adapted from Sorjonen et al. 2017). Right: Corresponding p-values of the approximate permutation test. 95.1% of the samples have a p-value > 0.05 (the samples above the horizontal line) indicating that the test can identify that the effect size is due to random chance.

NOTES

[i] We presume that X is on the horizontal axis and increases to the right, and Y is on the vertical axis and increases upwards.

[ii] The scope can be empirical (using the empirical minima and maxima of X and Y) or theoretical (using theoretical minimum and maximum of X and Y). The theoretical scope has a larger empty space relative to the scope than the empirical scope, hence results in a larger effect size. For most NCA applications we recommend using the empirical scope to avoid over-estimation of the effect size.

[iii] Standard bootstrap samples are obtained by sampling cases with replacement until the size of the bootstrap sample equals the size of the original sample. The standard deviation of the bootstrap sampling distribution can be used to estimate the standard error and the confidence interval. The confidence interval can be used for null hypothesis testing: testing whether the null is covered by the 95% confidence interval, which corresponds to null hypothesis testing with p-value 0.05. Producing a valid confidence interval for NCA's effect size implies that the estimated effect size has an upper confidence bound and a lower confidence bound. However, in NCA all cases are on or below the ceiling line by definition, and resampling of cases does not result in a ceiling line above the original ceiling line. Consequently, the lower bound effect size cannot be validly produced and bootstrapping will disproportionally often produce a (nearly same) effect size as the original effect size. Applying standard bootstrapping to NCA would result in invalid significance tests where the null hypothesis is excessively rejected.

[iv] In this example with three cases, the p-value cannot be smaller than 0.17.

[v] To illustrate this problem, it takes about one minute to calculate a p-value for 100,000 permutations on a personal computer. For N=10 the number of permutations is 3,628,800 and the

time for calculating a p-value is more than 30 minutes. The computation time for N=11 is more than six hours, for N=12 more than three days, for N=13 more than a month, for N=14 more than a year, for N=15 more than 20 years, and for N=16 more than a lifetime.

[vi] Other names for the approximate permutation test are 'Monte Carlo permutation test', and 'random permutation test'.

[vii] Several ceiling techniques can be selected within NCA. The two default ceiling techniques are Ceiling Envelopment - Free Disposal Hull (CE-FDH), which results is a non-decreasing step function ceiling line that can be used when the data are discrete, and Ceiling Regression - Free Disposal Hull (CR-FDH), which is a trend line through the corners of the CE-FDH step function and can be used for (practically) continuous data. For more information on ceiling techniques see Dul (2016).

[viii] The p-value that is estimated by approximate permutation follows a binomial distribution. The formula for the standard error of a binomial distribution is:

SE = $\sqrt{p(1-p)/n}$, where p is the estimated p-value and n is the number of permutations. For a 95% confidence level the p-accuracy of the estimated p-value is: $Accuracy = 1.96 * \sqrt{p(1-p)/n}$. For a desired particular p-accuracy, the minimum number of permutations is: $N = \left[\left(1.96\sqrt{p(1-p)}\right)/Accuracy\right]^2$.

[ix] The analytical approach uses a formula with several assumptions, including assumptions about distributions for calculating the standard error of the NCA effect size. Currently, no formula is available for this purpose, and such formula cannot be easily derived. One main problem is the discontinuity of NCA effect size as a function of the distribution (the data generation process).