How redefining statistical significance can worsen the replication crisis

Cole Randall Williams*

November 13, 2018

Abstract

In response to the replication crisis in science, a group of prominent scholars has proposed redefining statistical significance by reducing the p-value significance threshold from 0.05 to 0.005. Rather than solving the replication problem, I show that lowering the significance threshold can *increase* the rate of false positives by creating a negative selection effect. Thus, redefining statistical significance may not be a silver bullet for solving the replication crisis.

1 Introduction

The low success rate of replication studies in the social (Open Science Collaboration, 2012; Camerer et al., 2016, 2018), biological (Bogdan et al., 2017), and medical (Prinz, Schlange and Asadullah, 2011) sciences has been seen as a threat to the credibility of the scientific enterprise. Responding to this crisis in replication, a large group of prominent scientists and statisticians

*University of Vienna, Vienna, Austria, cole.randall.williams@univie.ac.at. This article was born out of a conversation with Nikhil Addleman and benefited from comments by Jean-Paul Carvalho, Stergios Skaperdas, and Pat Testa.

has called for a reduction in the p-value significance threshold from its conventional level of 0.05 to 0.005 (Johnson, 2013; Benjamin et al., 2018). This higher evidential burden is supposed to lower the *false positive rate*—the rate at which claimed discoveries are in fact untrue—thereby improving reproducibility.

However, this argument ignores the ways in which studies can differ in their propensities to produce false positives. This turns out to be important. In this article, I show that lowering the significance threshold, as has been proposed, can *increase* the false positive rate through what I call "negative selection." *Negative selection* occurs when lowering the significance threshold leads to a greater proportion of significant outcomes being produced by studies with the highest false positive rates.

A clear example of negative selection is found by considering *methodological bias*, whereby, features in the design and execution of studies tend to produce significant findings when they should not have been produced (Ioannidis, 2005; Simmons, Nelson and Simonsohn, 2011; John, Loewenstein and Prelec, 2012; Gelman and Loken, 2014). In this case, an unintended consequence of lowering the significance threshold is that the proportion of significant outcomes that are the result of bias increases. As a result, either the reduction to the false positive rate will be mitigated or it will increase.

A number of authors have expressed their own concerns with the proposal (Gelman and Robert, 2014; Amrhein and Greenland, 2018; McShane et al., 2018; Matthews, 2018; Trafimow et al., 2018; Lakens et al., 2018; Stahl and Pickles, 2018; Perezgonzalez and Frías-Navarro, 2017). While these authors vary in their points of emphasis and dispute certain details, there are several common threads. Firstly, there is an expressed distrust in selecting a single "universal" significance threshold to arbitrate scientific discoveries. Recommended alternatives are to tailor the threshold for the setting at hand (Lakens et al., 2018) or to abolish the reliance on thresholds altogether (Gelman and Robert, 2014; Trafimow et al., 2018; McShane et al., 2018; Amrhein and Greenland, 2018). More specific to the proposal, it has been suggested that lowering α will exacerbate misinterpretations of p-values that currently

plague research and exaggerate the focus on single p-values at the expense of other evidential factors (Amrhein and Greenland, 2018; Trafimow et al., 2018).

2 Negative Selection

We begin with a simple illustration of negative selection with an adaptation of a classic model of bias in science (Ioannidis, 2005; Maniadis, Tufano and List, 2014). This illustration expands on the model presented in the proposal (Benjamin et al., 2018) to permit some studies to follow unsound research practices. The basic idea is that reducing the significance threshold makes it harder for a sound study to obtain significance, increasing the proportion of significant outcomes that are unsound, and thus driving up the false positive rate. This is followed by a general discussion of negative selection.

2.1 Example: Bias in Research

Consider a unit mass of independent studies.[1] In each study, a researcher conducts a hypothesis test between a pair of null (H_0) and alternative (H_1) hypotheses. Let $\phi \in (0, 1)$ denote the proportion of null hypotheses that are true. The outcome of a study is significant if the hypothesis test yields a p-value less than the significance threshold α. A *false positive* occurs when a hypothesis test yields a significant outcome when the null is in fact true and the *false positive rate* is equal to the number of false positives divided by the total number of significant outcomes. In the large population limit, this is

$$R(\alpha) = Pr\left(H_0|\text{signficant}, \alpha\right) = \frac{Pr\left(\text{significant}, H_0|\alpha\right)}{Pr\left(\text{significant}|\alpha\right)}.$$

[1] Assuming a unit mass of studies characterizes the large population limit. Alternatively, one can introduce the addendum to each statement "as the number of studies goes to infinity almost surely."

Methodological bias leads a positive fraction λ of studies to report a significant outcome even when it should not have been. Refer to these as *unsound* studies.[2] The remaining $1-\lambda$ of studies are *sound*, producing α type I and $\beta(\alpha)$ type II error rates at significance threshold α. The statistical power of a sound study $1-\beta(\alpha) = Pr(p \leq \alpha|H_1)$ decreases with a reduction in α. To ensure smaller p-values constitute stronger evidence against the null than larger p-values, assume the density of p-values to be strictly decreasing under the alternative. Under these conditions, the false positive rate can be expressed as

$$R(\alpha) = w(\alpha) R_u(\alpha) + (1 - w(\alpha)) R_s(\alpha)$$

where $R_u(\alpha)$ and $R_s(\alpha)$ are the false positive rates for unsound and sound studies and $w(\alpha) = \frac{\lambda}{\lambda+(1-\lambda)(\phi\alpha+(1-\phi)(1-\beta(\alpha)))}$ is the fraction of significant outcomes that are produced by an unsound study. The false positive rate is thus a weighted average of the false positive rates for the individual types, with the weights endogenously determined by α. The bias in unsound studies ensures that they produce a higher rate of false positives ($R_u(\alpha) > R_s(\alpha)$ for all $\alpha < 1$) and that this difference $R_u(\alpha) - R_s(\alpha)$ increases with a reduction in α.

Reducing the significance threshold produces two effects. Firstly, there is the desirable effect of reducing the false positive rate for sound studies. This is seen by writing

$$R_s(\alpha) = \frac{\phi\alpha}{\phi\alpha + (1-\phi)(1-\beta(\alpha))} = \left(1 + \frac{1-\phi}{\phi}\frac{1-\beta(\alpha)}{\alpha}\right)^{-1} \quad (1)$$

and noting that, as the density of p-values under the alternative is strictly decreasing, the ratio $\frac{1-\beta(\alpha)}{\alpha}$ increases with a reduction in α.

The second effect is *negative selection*: reducing α increases the propor-

[2] A more straightforward interpretation of this example is that all studies have bias: an insignificant outcome is misreported as significant with probability λ, as in Ioannidis (2005) and Maniadis, Tufano and List (2014). However, the interpretation we follow is useful for characterizing negative selection.

tion of significant outcomes that are unsound $w(\alpha)$. Because unsound studies produce a higher rate of false positives $R_u(\alpha) > R_s(\alpha)$, negative selection increases the overall false positive rate. This immediately implies that ignoring negative selection by treating $w(\alpha)$ as fixed exaggerates the reduction in the overall false positive rate. Furthermore, when the difference between $R_u(\alpha)$ and $R_s(\alpha)$ is large enough (when α is small), negative selection dominates the first effect leading to an increase in the overall false positive rate. We formally state this finding in the following proposition which is illustrated by figure 1.

Proposition 1. *For any $\lambda \in (0,1)$, there exists $\alpha^* \in (0,1)$ such that lowering the significance threshold α increases the false positive rate if $\alpha < \alpha^*$.*

*Proof.*See appendix. ■

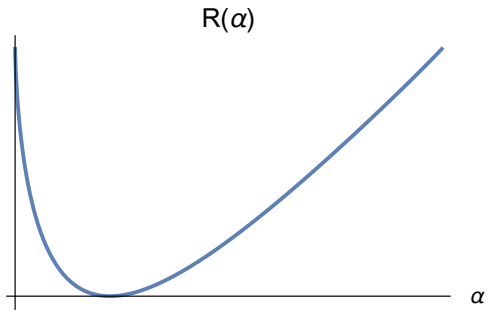

Figure 1: False positive rate $R(\alpha)$ as a function of the significance threshold α when there is positive bias $\lambda > 0$.

Thus, in the presence of negative selection, either redefining statistical significance will be harmful or its benefits mitigated. It is worth noting that this analysis implicitly assumes that reducing the significance threshold is not accompanied by efforts to increase statistical power (e.g. increasing sample sizes), for which the proponents of the policy also advocated. In considering implementing both measures, this result suggests that ignoring negative selection will overstate their combined effect. Moreover, a larger benefit might be seen by efforts to increase power alone.

2.2 Discussion

While bias in research is a particularly clear and empirically relevant source of negative selection, the effect should be understood as a general property of the false positive rate when there is heterogeneity among studies. Consider the case where there are many *types* of studies (indexed by t) that vary in prior chances, statistical power, researcher preferences, or capacity to exercise degrees of freedom. Again, by assuming the large population limit, the false positive rate can be conveniently expressed as

$$R(\alpha) = \sum_t w_t(\alpha) R_t(\alpha)$$

where $R_t(\alpha) = Pr(H_0|\text{signficant}, \alpha, t)$ is t's false positive rate and $w_t(\alpha) = Pr\left(t|\text{significant}, \alpha\right)$ is the fraction of significant outcomes that are of type t. It is clear from this expression that even if reducing α decreases each $R_t(\alpha)$, the change in $R(\alpha)$ will depend on $w_t(\alpha)$. If the same factors that lead certain studies to produce higher rates of false positives $R_{t'}(\alpha) > R_t(\alpha)$ also lead them to be less sensitive to changes in the significance threshold $w_{t'}\left(\alpha'\right) - w_{t'}\left(\alpha\right) > w_t\left(\alpha'\right) - w_t\left(\alpha\right)$, then negative selection will be present. See appendix A.2 for a general characterization of negative selection.

In closing, it is not clear what the path toward a more reliable science will entail, whether it be more stringent statistical requirements, increased adoption of Bayesian methods, further proliferation of preregistration, or even more radical changes. What is clear is that proposed measures to treat a problem of this import deserve rigorous scrutiny. A challenge for future work is to carefully analyze the various proposed actions in light of the complexities of actual research. Solutions that fail to do this are likely to be less effective or, as we have seen in this article, exacerbate the problem.

References

Amrhein, Valentin, and Sander Greenland. 2018. "Remove, rather

than redefine, statistical significance." *Nature Human Behaviour*.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E. J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P.A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. 2018. "Redefine statistical significance." *Nature Human Behaviour*.

Bogdan, Ryan, Betty Jo Salmeron, Caitlin E. Carey, Arpana Agrawal, Vince D. Calhoun, Hugh Garavan, Ahmad R. Hariri, Andreas Heinz, Matthew N. Hill, Andrew Holmes, Ned H. Kalin, and David Goldman. 2017. "Imaging Genetics and Genomics in Psychiatry: A Critical Review of Progress and Potential." *Biological Psychiatry*.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten,

Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating replicability of laboratory experiments in economics." *Science*, 351(6280): 1433–1436.

Camerer, Colin F, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, and Others. 2018. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour*, 1.

Gelman, A., and C. P. Robert. 2014. "Revised evidence for statistical standards." *Proceedings of the National Academy of Sciences*.

Gelman, Andrew, and Eric Loken. 2014. "The statistical Crisis in science." *American Scientist*, 102(6): 460–465.

Ioannidis, John P A. 2005. "Why most published research findings are false." *PLoS Medicine*, 2(8): 0696–0701.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science*.

Johnson, V. E. 2013. "Revised standards for statistical evidence." *Proceedings of the National Academy of Sciences*.

Lakens, Daniel, Federico G. Adolfi, Casper J. Albers, Farid Anvari, Matthew A.J. Apps, Shlomo E. Argamon, Thom Baguley, Raymond B. Becker, Stephen D. Benning, Daniel E. Bradford, Erin M. Buchanan, Aaron R. Caldwell, Ben Van Calster, Rickard Carlsson, Sau Chin Chen, Bryan Chung, Lincoln J. Colling, Gary S. Collins, Zander Crook, Emily S. Cross, Sameera Daniels, Henrik Danielsson, Lisa Debruine, Daniel J. Dunleavy, Brian D. Earp, Michele I. Feist, Jason D. Ferrell, James G.

Field, Nicholas W. Fox, Amanda Friesen, Caio Gomes, Monica Gonzalez-Marquez, James A. Grange, Andrew P. Grieve, Robert Guggenberger, James Grist, Anne Laura Van Harmelen, Fred Hasselman, Kevin D. Hochard, Mark R. Hoffarth, Nicholas P. Holmes, Michael Ingre, Peder M. Isager, Hanna K. Isotalus, Christer Johansson, Konrad Juszczyk, David A. Kenny, Ahmed A. Khalil, Barbara Konat, Junpeng Lao, Erik Gahner Larsen, Gerine M.A. Lodder, Jiří Lukavský, Christopher R. Madan, David Manheim, Stephen R. Martin, Andrea E. Martin, Deborah G. Mayo, Randy J. McCarthy, Kevin McConway, Colin McFarland, Amanda Q.X. Nio, Gustav Nilsonne, Cilene Lino De Oliveira, Jean Jacques Orban De Xivry, Sam Parsons, Gerit Pfuhl, Kimberly A. Quinn, John J. Sakon, S. Adil Saribay, Iris K. Schneider, Manojkumar Selvaraju, Zsuzsika Sjoerds, Samuel G. Smith, Tim Smits, Jeffrey R. Spies, Vishnu Sreekumar, Crystal N. Steltenpohl, Neil Stenhouse, Wojciech Świątkowski, Miguel A. Vadillo, Marcel A.L.M. Van Assen, Matt N. Williams, Samantha E. Williams, Donald R. Williams, Tal Yarkoni, Ignazio Ziano, and Rolf A. Zwaan. 2018. "Justify your alpha." *Nature Human Behaviour*.

Maniadis, Zacharias, Fabio Tufano, and John a. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *American Economic Review*.

Matthews, Robert A J. 2018. "Beyond 'significance': principles and practice of the Analysis of Credibility." *Royal Society Open Science*, 5(1).

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. 2018. "Abandon Statistical Significance."

Open Science Collaboration. 2012. "The reproducibility of psychological science." *Science*.

Perezgonzalez, Jose D., and M. Dolores Frías-Navarro. 2017. "Retract p < 0.005 and propose using JASP, instead." *F1000Research.*

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe it or not: How much can we rely on published data on potential drug targets?" *Nature Reviews Drug Discovery.*

Schervish, Mark J. 1997. *Theory of Statistics.* Springer Series in Statistics.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology." *Psychological Science*, 22(11): 1359–1366.

Stahl, D, and A Pickles. 2018. "Fact or fiction: reducing the proportion and impact of false positives." *Psychological medicine*, 48(7): 1084–1091.

Trafimow, David, Valentin Amrhein, Corson N. Areshenkoff, Carlos J. Barrera-Causil, Eric J. Beh, Yusuf K. Bilgiç, Roser Bono, Michael T. Bradley, William M. Briggs, Héctor A. Cepeda-Freyre, Sergio E. Chaigneau, Daniel R. Ciocca, Juan C. Correa, Denis Cousineau, Michiel R. de Boer, Subhra S. Dhar, Igor Dolgov, Juana Gómez-Benito, Marian Grendar, James W. Grice, Martin E. Guerrero-Gimenez, Andrés Gutiérrez, Tania B. Huedo-Medina, Klaus Jaffe, Armina Janyan, Ali Karimnezhad, Fränzi Korner-Nievergelt, Koji Kosugi, Martin Lachmair, Rubén D. Ledesma, Roberto Limongi, Marco T. Liuzza, Rosaria Lombardo, Michael J. Marks, Gunther Meinlschmidt, Ladislas Nalborczyk, Hung T. Nguyen, Raydonal Ospina, Jose D. Perezgonzalez, Roland Pfister, Juan J. Rahona, David A. Rodríguez-Medina, Xavier Romão, Susana Ruiz-Fernández, Isabel Suarez, Marion Tegethoff, Mauricio Tejo, Rens van de Schoot, Ivan I. Vankov, Santiago Velasco-Forero, Tonghui Wang, Yuki Yamada, Felipe C.M. Zoppino, and Fernando Marmolejo-Ramos. 2018. "Manipulating the alpha level cannot cure significance testing." *Frontiers in Psychology.*

A Appendix

A.1 Proofs for Main Text

Begin by verifying that the ratio $\frac{1-\beta(\alpha)}{\alpha}$ is decreasing in α. By definition, p-values are uniformly distributed between 0 and 1 under the null. Let $1-\beta(\alpha) = Pr(p \leq \alpha | H_1)$ be the CDF of p-values under the alternative so that $(1-\beta(\alpha))'$ is the density. By assumption, the density of p-values under the alternative hypothesis is decreasing $(1-\beta(\tilde{\alpha}))' > (1-\beta(\alpha))'$ for $\tilde{\alpha} < \alpha$. An implication of this assumption is

$$1-\beta(\alpha) = \int_0^\alpha (1-\beta(p))' \, dp > \int_0^\alpha (1-\beta(\alpha))' \, dp = (1-\beta(\alpha))' \alpha$$

which further implies that the ratio $\frac{1-\beta(\alpha)}{\alpha}$ is decreasing in α

$$\left(\frac{1-\beta(\alpha)}{\alpha}\right)' = \frac{(1-\beta(\alpha))'\alpha - (1-\beta(\alpha))}{\alpha^2} < 0.$$

Proof of Proposition 1. We want to show that for $\lambda \in (0,1)$, $R(\alpha)$ is decreasing up to some point α^* and increasing thereafter. Differentiating $R(\alpha) = \left(1 + \frac{1-\phi}{\phi} \cdot \frac{\lambda+(1-\lambda)(1-\beta(\alpha))}{\lambda+(1-\lambda)\alpha}\right)^{-1}$ yields $R'(\alpha)$ with the same sign as

$$h(\alpha) \equiv -\lambda\left((1-\beta(\alpha))' - 1\right) - (1-\lambda)\left(\alpha(1-\beta(\alpha))' - (1-\beta(\alpha))\right).$$

Upon differentiating, $h'(\alpha) = -(\lambda + (1-\lambda)\alpha)(1-\beta(\alpha))'' \geq 0$. Noting that $\lim_{\alpha \to 0^+} h(\alpha) < 0$ and $\lim_{\alpha \to 1^-} h(\alpha) > 0$ whenever $\lambda \in (0,1)$ completes the proof. ■

A.2 General Characterization of Negative Selection

This section provides a general characterization of negative selection. Suppose there are many different *types* of studies $t \in T$ that vary in prior chances, statistical power, researcher preferences, or capacity to exercise degrees of

freedom. Types are distributed according to probability measure μ with probability space (T, Σ, μ). Continue assuming that the population is comprised of a unit mass of studies. For the false positive rate to be well-defined, we require the mapping $t \mapsto Pr\left(\text{significant}, H_\theta|\alpha, t\right)$ to be measurable for all $\alpha \in [0,1]$ and $\theta \in \{0,1\}$. The false positive rate of each type of study

$$R_t(\alpha) = \frac{Pr\left(\text{significant}, H_0|\alpha, t\right)}{Pr\left(\text{significant}|\alpha, t\right)}$$

is assumed to be nondecreasing in α for all $t \in T$.[3] Let $\eta(D|\alpha)$ be the proportion of significant outcomes that are of a type in $D \in \Sigma$ at significance threshold α, and write

$$R(\alpha) = \int R_t(\alpha) \mathrm{d}\eta(t|\alpha).$$

For the formal definition of negative selection, define $B_{\delta,\alpha} = \{t \in T | R_t(\alpha) > \delta\}$ so that $(B_{\delta,\alpha}, B_{\delta,\alpha}^C)$ partitions types between those with the highest false positive rates $B_{\delta,\alpha}$ and those with the lowest false positive rates $B_{\delta,\alpha}^C$.

Definition. *Negative selection* occurs when lowering the significance threshold from α to α' increases the proportion of significant outcomes that are of types with the highest false positive rates, that is, $\eta(B_{\delta,\alpha'}|\alpha') > \eta(B_{\delta,\alpha}|\alpha)$ for some $\delta \in (0,1)$.

We now show that reducing the significance threshold increases the false positive rate if and only if negative selection is strong enough to counteract the decrease in the false positive rates for the individual types $R_t(\alpha)$. The condition for when this occurs will now be given, followed by a formal statement of the result.

Condition 1.

$$\eta(B_{\delta,\alpha'}|\alpha') - \eta(B_{\delta,\alpha}|\alpha) > \psi(\delta, \alpha, \alpha') > 0 \tag{2}$$

The precise form of $\psi(\delta, \alpha, \alpha')$ is given in the proof that follows.

[3] Absent this assumption, the potential for lowering the significance threshold to increase the false positive rate is trivial.

Theorem 1. Reducing the significance threshold *increases* the false positive rate if and only if *negative selection* satisfies condition 2.

Proof of Theorem 1. The measurability of $B_{\delta,\alpha}$ is a consequence of the measurability of $t \mapsto Pr\left(\text{significant}, H_\theta|\alpha, t\right)$. Denote the conditional expectations $r(\delta,\alpha) \equiv \mathbb{E}[R_t|B_{\delta,\alpha},\alpha]$ and $s(\delta,\alpha) \equiv \mathbb{E}[R_t|B^C_{\delta,\alpha},\alpha]$. Making the condition in (1) explicit

$$\psi(\delta,\alpha,\alpha') = \left(\frac{r(\delta,\alpha)-s(\delta,\alpha)}{r(\delta,\alpha')-s(\delta,\alpha')}-1\right)\eta(B_{\delta,\alpha}|\alpha)+\frac{s(\delta,\alpha)-s(\delta,\alpha')}{r(\delta,\alpha')-s(\delta,\alpha')}. \quad (3)$$

Expand the false positive rate

$$R(\alpha) = r(\delta,\alpha)\eta(B_{\delta,\alpha}|\alpha)+s(\delta,\alpha)\eta(B^C_{\delta,\alpha}|\alpha). \quad (4)$$

By (3) and (4)

$$\operatorname{sgn}\left(R\left(\alpha'\right)-R\left(\alpha\right)\right) = \operatorname{sgn}\left(\eta\left(B_{\delta,\alpha'}|\alpha'\right)-\eta(B_{\delta,\alpha}|\alpha)-\psi\left(\delta,\alpha,\alpha'\right)\right). \quad (5)$$

From (5), it follows that if condition 2 holds, then $R(\alpha') > R(\alpha)$. Going the other direction, (5) shows that if $R(\alpha') > R(\alpha)$ then the first inequality in condition 2 must hold. All that remains to be shown is, if $R(\alpha') > R(\alpha)$, then the second inequality in condition 2 holds. That is, there exists δ satisfying $\psi(\delta,\alpha,\alpha') > 0$.

To obtain a contradiction, assume $\psi(\delta,\alpha,\alpha') \leq 0$ and thus $\eta(B_{\delta,\alpha'}|\alpha') \leq \eta(B_{\delta,\alpha}|\alpha)$ for all δ. Define $\nu(\cdot|\alpha)$ to be the measure induced on $[0,1]$ by the mapping $R_t(\alpha) \mapsto \delta$ such that $R_t(\alpha) = \delta$. By assumption

$$\nu\left([0,\delta]\,|\alpha'\right) = \eta\left(B^C_{\delta,\alpha'}|\alpha'\right) \geq \eta\left(B^C_{\delta,\alpha}|\alpha\right) = \nu\left([0,\delta]\,|\alpha\right)$$

for all δ, and thus $\nu(\cdot|\alpha)$ has first-order stochastic dominance over $\nu(\cdot|\alpha')$. By theorem A.81 in Schervish (1997)

$$R(\alpha) = \int R_t(\alpha)\mathrm{d}\eta(t|\alpha) = \int_0^1 \delta\mathrm{d}\nu(\delta|\alpha).$$

By consequence of first-order stochastic dominance, $\int_0^1 \delta \mathrm{d}\nu(\delta|\alpha) \geq \int_0^1 \delta \mathrm{d}\nu(\delta|\alpha')$ and thus $R(\alpha) \geq R(\alpha')$ contradicting $R(\alpha) < R(\alpha')$. ■