Replication of Study The nonsense math effect by Kimmo Eriksson (2012, Journal of Judgment and Decision Making)

Dr. Christine Blech christine.blech@fernuni-hagen.de

Nadja Bahr
FernUniversität in Hagen
(nadjasamia87@gmail.com)

Ariella Sonder
FernUniversität in Hagen
(ariella.sonder@studium.fernuni-hagen.de)

Introduction

Original citation

Eriksson, K. (2012). The nonsense math effect. *Judgment and Decision Making, 7*, 746–749.

Original Study

Eriksson (2012) examined in his study the *nonsense math effect* the tendency of a positive bias for meaningless math in judgments of quality research. Furthermore, Eriksson analysed whether there are differences between academic fields and their ability to critically evaluate meaningless mathematics in scientific abstracts, since "in those disciplines where most researchers do not master mathematics, the use of mathematics may be held in too much awe" (Eriksson, 2012, p. 746). Two abstracts were given to judge. Either the first or the second abstract was randomly manipulated with an added math formula to control that the effect was not caused by one specific abstract. The author did provide evidence that an abstract with an added, but meaningless math sentence will be rated higher than an abstract without math. However, the effect was not found among participants with degrees in mathematics, natural science, technology or medicine. Participants did rate higher, if there is a senseless math formula included, no matter in which abstract. The evidence was confirmed in a Two Way ANOVA, $F(1,192) = 8.68$, $p < .05$ (Eriksson, 2012, p. 748). However, the author noted that presumably lack of math skills in the fields outside math, science and technology decrease critical evaluation of math. Since this was not tested in the experiment, the evidence should be interpreted with caution. The author suggested to examine whether training in math could decrease this bias.

Target of Replication

In the current replication of Eriksson (2012) we seek to reproduce the methodology used in the original paper. However, referring to the requirements of the Hagen Cumulative Science Project I, we focussed on one key effect, the effect of the area of degree on the rating advantage of added math. The aim of the study was to show that there is a rating advantage in abstracts with an added formula even though the formula is not related to the topic. The author's principal research questions were whether there is a rating advantage of added math, regardless of which abstract is manipulated, and whether the rating advantage is moderated by participant`s area of degree. Accordingly, the author reported in the one sample t-test a significant positive rating of the abstracts with added math, *M* = 4.7, *SD* = 21.0, *n* = 200, *p* < .01 (Eriksson, 2012, p. 748). However, the effect was not found in the areas of degree in math, technology and science, *M* = -1.3, *SD* = 19.2, *n* = 69, *p* > .05 and medicine, *M* = 3.0, *SD* = 16.0, *n* = 16, *p* > .05 (Eriksson, 2012, p. 748). Additionally, binomial tests were done, the results were that the majorities for humanities and social science and other (e.g., education) were found to be statistically significant, *p* < .05 (Eriksson, 2012, p. 748). Furthermore, the ANOVA confirmed the result with a main effect of area of degree, *F*(3,192) = 4.2, *p* < .01 (Eriksson, 2012, p. 748). A successful replication would hence find an effect on higher rating of the abstract with an added math formula by participants outside the academic fields of mathematics, technology and science. According to the original paper we plan the study among German speaking participants with bachelor`s, master`s degree or PhD from different areas of math, science and technology as well as social science, humanities and others.

Method

Power Analysis

The necessary sample size for the effect size was calculated with an a priori power analysis using G*Power (Faul, Erdfelder, Lang & Buchner, 2009). Since the effect size was not reported in original study, partial eta squared (η^2_p) was additionally computed in reanalysis of original data applied in SPSS. Using G*Power the effect size η^2_p was calculated into the asked effect size f needed for test family "F-test" and "ANOVA: Fixed effects, special, main effects and interactions" as the statistical test. We chose this test, because we compared means of multiple independent groups defined by multiple categorical independent variables. With the error probability α = .05 a sample size of *N* = 170 is required to achieve a power of .80, *N* = 191 will be needed for a power of .85, *N* = 219 for a power of .90 and *N* = 264 for a .95 power. In order to achieve a minimum power of .8 we aim to achieve a minimum sample size of 170 participants. However, we aim to reach 192 participants to have a test power of .85. Thus, we will include also participants with a bachelor´s degree.

Planned Sample

Based on the power analysis mentioned above, we plan to include 191+1 participants in the analysis in order to achieve a power of .85 to keep number of participants in both conditions even. In case we achieve more participants during our time schedule, we will include participants until a maximum of *N* = 264. This sample size comes with a power of .95. To reach a preferably higher number of participants, people with a bachelor`s degree will be included**,** presumed they have "experience of reading research reports, such as journal articles, conference papers, anthology chapters or monograph" (Eriksson, 2012, p. 747) as it is part of their studies. Dependent on the number of participants and the amount of participants with master`s and PhD degrees, participants with bachelor`s degree will be excluded from the main analyses and used as a separate category in further exploratory data analyses. The study will be administered as an online-questionnaire in the "virtual laboratory" of the FernUniversität Hagen. Therefore, a sample consisting of students from the FernUniversität Hagen, Germany is expected. In order to get a preferably similar group of participants as in the original study, academics will be contacted through email, social media, scientific forums, Alumni (networks or groups of former students) and academic associations in Austria, Germany and Switzerland. Participants can win one out of three 50 Euros vouchers. Students in master`s programme, who fit the sample and request them, will get credit points in return to participate.

Materials

Sociodemographic data

According to the original study, participant`s age and gender will be also collected.

Qualifications

The study starts with a questionnaire about participant`s qualification. Data about the degree, the area of degree and the experience of reading scientific papers will be collected. Participants can choose between either "*Have read less than 10 different reports*", "*Have read between 10 and 100 different reports*", or "*Have read more than 100 different reports*". This corresponds to the descriptions of the original article with the following adjustments: The questionnaire will contain the option "bachelor`s degree" according to the possible inclusion of bachelor graduates. The area of degree will be surveyed in an open question format and later on classified similar to the original study into four broad areas of research: Humanities or social science, medicine, mathematics, natural science or technology or other (e.g., education, economics).

Scientific abstracts

"The two abstracts were taken from real research papers, well-cited and published in very good journals. They were selected so that they would be generally understandable to non-specialists" (Eriksson, 2012, p. 747). The abstracts were selected as they were used in the original study. After we translated them into German, they were lectured by an active researcher with a PhD in Sociology to ensure quality.

First Abstract: "Foraging" - Risk and reciprocity in Meriam food sharing

The first abstract is "a study of whether foodsharing practices among a foraging tribe could be predicted from risk reduction and reciprocity (from Bliege Bird et al. 2002)" (Eriksson, 2012, p. 747).

Second Abstract: "Incarceration" – The mark of a criminal record

The second abstract is "a study of the consequences of incarceration for the employment outcomes of black and white job seekers (from Pager 2003)" (Eriksson, 2012, p. 747).

A mathematical model

The mathematical Model TPP=T0−fT0df2−fTPdf (Soetens et al., 1984 as cited in Eriksson 2012) is taken from an unrelated study about reactions times in choice experiments. Neither of the abstracts is about reaction times or sequential effects, thus the inclusion of the math Model is meaningless. The manipulation consists in the random addition of this sentence at the end of one of the two different abstracts in two conditions: Condition 1- abstract 1 is added with the formula, condition 2- abstract 2 was added with the formula.

Procedure

"Participants filled in an online questionnaire. It started with questions about their qualifications, including their postgraduate degree [either *Master's degree* (88%) or *PhD* (12%)]; the area of their degree [either *humanities or social science* (42%), *medicine* (8%), *mathematics, natural science or technology* (34%), or *other (e.g., education)* (16%)]; and their experience of reading research reports, such as journal articles, conference papers, anthology chapters or monographs [either *Have read less than 10 different reports* (12%), *Have read between 10 and 100 different reports* (54%), or *Have read more than 100 different reports* (34%)]. The questionnaire went on to describe that organizers of scientific conferences often ask for researchers to submit abstracts of the research they would like to present, and that based only on these short abstracts the highest quality research is to be selected. The current study was presented as an investigation of how readers of abstracts judge the quality of research.

Two abstracts were then presented. For each abstract, participants were asked to give their general judgment of quality. The study will be presented as an examination about how readers judge the quality of scientific abstracts. To mimic the actual examination, there will be also questions about other aspects of the abstracts, e.g. how important or interesting the abstracts were" (Eriksson, 2012, p. 747). This was followed precisely. Like in the original study, the current study will be distributed online (via the "virtually laboratory" of the FernUniversität Hagen). The questionnaire comes in two versions, either the first abstract is manipulated or the second one. Responses were given on a scale from 0 (*the very lowest quality*) to 100 (*the very highest quality*).

To mimic a typical procedure for judgment of submitted abstracts, participants were also asked to rate some other aspects of the abstract, such as the importance and how interesting it was. At the end of the study, participants get the possibility to give their email address in case they want to join the lottery to win the amount of 50 euros or get a credit point. Furthermore, we inform the participants to get the results in case they are interested.

Following the original study, we compute the 2 abstracts to be rated (Foraging x Incarceration) x 2 conditions (math added to Foraging or to Incarceration) x 4 areas of degree (math/science/technology, social science/humanities, medicine or other (e.g., education). First the dependent variable *rating advantage of added math* is calculated as the rating of the manipulated abstract minus the rating of the non-manipulated abstract. After that we split the sample by the factor area of degree and compute descriptive statistics and a one sample t-Test. In the next step we analyze the data through a univariate Two Way ANOVA with rating advantage as the dependent variable, with condition and area of degree as factors. After that, we add the covariate reading experience to the ANOVA. Finally, as in the original study, we examined the effect of the manipulation in terms of percentage who rated the abstract with added math highest with binomial tests.

Differences from the original study

The sample of the study might contain a majority of German students while in the original study by Eriksson (2012) the sample consisted of U.S citizens recruited on the job platform "Amazon's Mechanical Turk" in return for pay. However, we will aim to have a preferable similar sample size with approximately 90 percent of master`s degree and about 10% PhD, with about 50 percent from the math/technology/sciences field or medicine and 50 percent from the academic disciplines of social sciences and others. This could still lead to differences in our survey, e.g. we expect that the original sample consisted of participants of a wider range in education. Additionally, the translation of the abstracts might affect participant`s judgments about the quality. Furthermore, since we only have a short time frame for the empirical survey, we choose to also include participants with a bachelor`s degree. After the survey we will analyse the data and exclude people with bachelor`s degree from the main analyses in case we have reached a sufficient number of approximately 200 participants with the given characteristics mentioned above. To differentiate the group differences closer, we will also compute post-hoc tests or contrast analysis.

(Post Data Collection) Methods Addendum

Actual Sample

For our replication study, we recruited 289 participants via the virtual laboratory of the FernUniversität Hagen, the University research pool, Facebook and through "snowball effect" in Germany, Switzerland and Austria. The data collection time frame was from 09.08.2019 until 01.09.2019. We excluded four participants because of a too short time duration, two participants denied the data protection agreement and another five did not have a bachelor`s degree at least. Furthermore, we excluded 14 more participants who did not participate seriously. Of the remaining 264 participants 56.4% were female. The mean age of the participants was 36.68 years (*SD* = 11.51). The sample consists of 45.5% with a master`s degree, 31.1% with a bachelor`s degree, another 20% have a PhD and 3.4% have another degree (e.g. German degree "Staatsexamen" which is equivalent to at least master`s degree level). Compared to the planned sample, we did achieve the desired sample size of 264 participants to ensure a test power of 95%.

Results

The mean value for the rating advantage in the one sample t-test was *M* = 4.39 (*SD* = 26.33) in the full sample. Table 1 shows descriptive statistics and the results of the one-sample t-test in comparison to the results of the reanalysis. The table indicates, as in the original study, an overall positive effect of added math, *t*(263) = 2.71, *p* = .007, *d* = 0.33, but unlike the original study, no impact of the area of degree on the rating advantage of added math. The means between the subgroups divided by area of degree did not differ significantly.

Table 1

Descriptive Statistics comparison and t-test results

Study		Reanalysis of Original			Replication	
Area of degree	*N*	*M*	*SD*	*N*	*M*	*SD*
Math, Science, Technology	69	-1.28	(19.24	68	3.47	(22.92)
Medicine	16	3.06	(15.99)	18	8.44	(20.57)
Humanities, Social Science	84	**6.60****	(21.15)	136	3.56	(28.86)
Others, e.g. education	31	**13.90****	(23.31)	42	6.86	(25.58)
Total	200	**4.74****	(21.01)	264	**4.39****	(26.33)

Notes. ** p < .01, * p < .05.

To replicate the primary finding, we conducted an univariate Two Way ANOVA with *rating advantage of added math* as dependent variable and *condition* and *area of degree* as factors. The ANOVA confirmed, that there was a rating advantage of added math*, F*(1,256) = 5.96, *p* = .02, η_{p^2} = .023. However, there was no main effect of *area of degree*, *F*(3,256) = 0.20, *p* =.89, η_{p^2} = .002, meaning that the rating advantage of added math did not differed between participants in different areas. Thus, the primary finding of the original study was not replicated. There was a main effect of condition, *F*(1, 256) = 7.29, *p* = .007, $\mathrm{\eta_{p^2}}$ =.028, meaning, that regarding of which abstract was manipulated, the rating advantage of added math differed. Like in the original study, there was no significant interaction between *condition* and *area of degree*, *F*(3, 256) = 0.90, *p* = .44, η_{p^2} =.010. The results of the ANOVA are robust to inclusion of reading experience of research reports as a covariate. The covariate reading experience was significant, *F*(1, 255) = 15.56, *p* < .001, η_{p^2} = .058. Figure 1 shows a comparison of the original study and the replication. It illustrates the effect of the manipulation in terms of the percentage who rated the abstract with added math higher. As in the original study, we excluded those who gave equal ratings of both abstracts. The majorities for “social science” and “other” were significant in the original study, *p* < .05, computed with a binomial test. In our replication, no group was significant. That means, the majorities did not differ significantly. As seen in figure 1, in the replication, the group differences are smaller than in the original study.

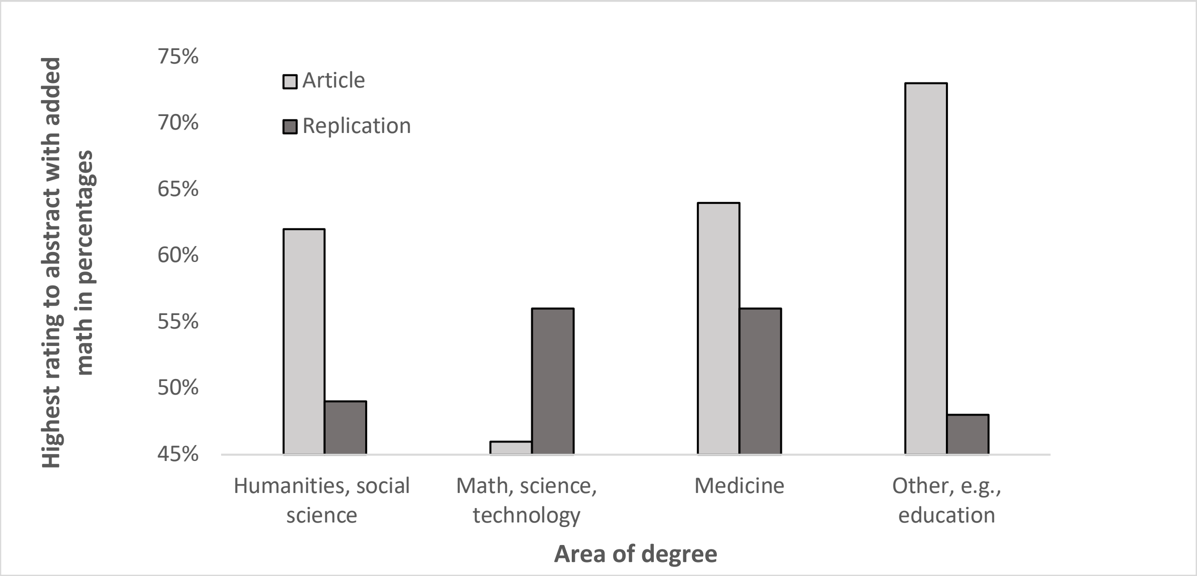

Figure 1. comparison of original and replication, percentages, who gave the highest rating to the abstract with added math.

Exploratory analyses

Thus, the factor *area of degree* did not have an impact on the dependent variable rating *advantage of added math,* we examined the data referring the categorization and found discrepancies between the categorization into the area of degree and the specific subject of degree. For example, participants with a degree in business, philosophy or law did categorize themselves in either the *area humanities* and *social science* or into *others*. It was not possible to recategorize the data with the system in the original study was used, because the category "*others, e.g. education*" was not specified further. Thus, we recategorized with the German, official *DFG Fachsystematik* (Deutsche Forschungsgesellschaft, 2017) and found different results. Referring to the *DFG Fachsystematik*, there are 4 big science areas: Social sciences, Life sciences, Engineering sciences and Natural sciences. There was an effect with an one-sample t-test for the area *life sciences,* which contains medicine (*M* = 8.47, *SD* = 27.56, *n* = 34), on the rating advantage of added math, *t*(33) = 2.48, *p* = .02, *d* = 0.43. Further, as for the group of *social sciences (M* = 4.03*, SD* = 27.56*, n* = 183) an equivocal result of *t*(182) = 1.98, *p* = .05, *d* = 0.15 was found. An univariate, two factor ANOVA could not detect any effect of the area of degree on the rating advantage of added math, *F*(3,256) = .41, *p* = .75, η_{p^2} =.005.

Moreover, since the factor *condition* was significant, we examined descriptive statistics for the means of the ratings of the abstracts for both conditions. In an independent samples t-test, the abstract "Foraging" got higher ratings when the math formula was added (*M* = 57.55*, SD* = 21.83), than in the non-manipulated version *(M* = 51.18, *SD* = 21.18). This is a significant mean difference, according to an independent samples t-test, *t*(262) = 2.41, *p* =.02, *d* = 0.30. According to Cohen (1988) this is a small effect. The abstract "Incarceration", regardless of the added math formula *(M* = 58.20, *SD* = 20.71 in condition *1, without math, M*= 60.46, *SD* = 24.44 in condition 2, with math) got no significantly higher rating with added math, *t*(262)= -.81, *p* = .42, *d* = 0.10. However, the abstract "Incarceration" got higher ratings for both conditions than abstract "Foraging". On the contrary in the original study, the abstract with the added formula got the higher rating.

Finally, since there was a sufficiently large sample of *N* = 182 without the bachelor graduates, once again we examined with only master graduates and participants with a PhD, because the original study`s sample did not contain participants with a bachelor´s degree. In a one-sample t-test, no significant overall rating advantage of added math was found, *M* = 2.92, *SD* = 24.86*, n* = 182, *t*(181) = 1.58, *p* = .12, *d* = 0.24. No significant mean differences were found in a one-sample t-test divided into the groups of area of degree. An univariate two-factorial ANOVA confirms these results with *F*(1,174) = 2.28, *p* = .13, η_{p^2} = .013.

Discussion

Summary of Replication Attempt

Our results reveal that the replication of the key effect (effect of the area of degree on the rating advantage of added math) failed. However, an overall positive effect of added math could be found and was confirmed through the ANOVA. However, the effect is not found when excluding the bachelor`s degree participants. On the contrary to the original study, we found an effect of the factor *condition* on the *rating advantage of added math*, meaning that regarding of which abstract was manipulated, the manipulation affects differences in the rating advantages. Moreover, we examined the mean differences of the rating of the abstracts in both conditions with t-tests for independent samples. Abstract two was rated higher in both conditions and the rating advantage of added math was significant only for abstract one. This is to say, there was a rating advantage. However, the condition moderated the rating advantage of added math, whereas, unlike the original study, the area of degree did not have an impact. Possible reasons for this result will be discussed in the commentary section.

Commentary

Three possible reasons might explain differences in the key effects in the original and replication study and are therefore examined more closely:

Comparing the samples:

In the original study the sample was from the USA, in the replication from Germany, Austria and Switzerland. Thus, cultural differences in perception and quality ratings could have had an impact on the results. Further, in the original sample, 88% of the participants had a master`s degree while in the replication study the sample consisted only of 68.9% participants with master`s degree or higher (another 19.7% with a PhD and 3.8 % with another degree, e.g. German degree "Staatsexamen" which is equivalent to master`s degree level), 31.1 % had a bachelor`s degree.

Moreover, the different education systems have to be taken into consideration. As suggested in the original study, the possible correlation between math education and holding math in wrong awe should be examined further.

Table 2:

Comparison of the samples

study	*original study*	*replication study*
Recruitment method	Mechanical Turk	University pool, Facebook Groups, Friends
Country	USA	Germany, Austria, Switzerland
M Age	31.58 (9.21)	36.68 (11.51)
Sex (m/w)	54.0%/46.0%	43.6%/56.4%
Master`s/PhD	100%	68.9%

Differences in the study materials:

Differences in the study materials could occur due to the translation into the German. In our online survey, we gave participants the possibility to let us know their opinion about the study. The opinions were clearly in one direction. The abstracts had substantial differences of quality and were not comparable. The second abstract "Incarceration" was rated higher by several opinions. This opinion was also confirmed by the descriptive statistics of the mean ratings of both abstracts. The mean quality rating of the second abstract was higher in both conditions (with and without added math), than the mean quality rating of the first abstract. Thus, we conclude that the two abstracts differed in quality. Further, we conclude that not only the added math but also the abstract quality (abstract topic, relevance for society, personal preference) has an influence on the rating advantage of added math. Thus, the internal validity might be affected negatively, since the abstract quality itself was measured in combination with the math formula and not separated- only the math formula- as it was assumed in the original study.

Categorization method problem

Furthermore, different categorization methods may lead to different results. There was no clear categorization system. Participants with the same subject of degree did categorize themselves into two different groups, moreover, it was not possible to recategorize with these 4 groups since the last category "*others, e.g. education*" was not defined clearly, what subject could be else belong to the category.

References:

1. Bliege Bird, R., Bird, D. W., Smith, E. A., & Kushnick, G. C. (2002). Risk and reciprocity in Meriam food sharing. *Evolution and Human Behavior*, *23*(4), 297–321. https://doi.org/10.1016/S1090-5138(02)00098-3
2. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J: L. Erlbaum Associates.
3. Deutsche Forschungsgesellschaft. (2017). www.dfg.de. Abgerufen 26. September 2019, von https://www.dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/amtsperiode_2016_2019/fachsystematik_2016-2019_de_grafik.pdf.
4. Eriksson, K. (2012). The Nonsense Math Effect. *Judgment and Decision Making 7,* 746–749.
5. Pager, D. (2003). The Mark of a Criminal Record. *American Journal of Sociology*, *108*(5), 937–975. https://doi.org/10.1086/374403

6. Soetens, E., Deboeck, M., & Hueting, J. (1984). Automatic aftereffects in two-choice reaction time: A mathematical representation of some concepts. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(4), 581–598.