Meta-Psychology, 2024, vol 8, MP.2020.2639
https://doi.org/10.15626/MP.2020.2639
Article type: File-Drawer Report

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Alex O. Holcombe
Reviewed by: Matthew Inglis, Tom Wallis
Analysis reproduced by: Lucija Batinović
All supplementary files can be accessed:
https://doi.org/10.17605/OSF.IO/75RFZ/

The Effect of Variety on Perceived Quantity: Failures to Replicate Redden and Hoch (2009)

Lukas Röseler[1,2,4], Georg Felser[2], Jana Asberger[3], and Astrid Schütz[1]
[1]Department of Personality Psychology and Psychological Assessment, University of Bamberg, Bamberg, Germany
[2]Business Psychology, Harz University of Applied Sciences, Wernigerode, Germany
[3]Educational Research and Methodology, University of Erfurt, Erfurt, Germany
[4]Münster Center for Open Science, University of Münster, Münster, Germany

Redden and Hoch (2009) found that variety in a set of items robustly decreased the perceived quantity of the sum of these items across multiple studies. For example, a set of multicolored M&M's was estimated to contain fewer M&M's than an equally large set of single-colored M&M's (e.g., Redden and Hoch, 2009, Study 3). We conducted six studies with methods that were similar to those used by Redden and Hoch (2009) and did not find this effect in any of them. A meta-analysis of the four original studies and 6 replication studies (N = 1,383) revealed no evidence for the phenomenon that variety reduces perceived quantity.

Keywords: file-drawer report, quantity estimation, variety, Gestalt, replication

Transparency Statement

This report is an exhaustive report on all data available from research projects related to the estimation of variety and perceived quantity of items in which GF was the principal investigator. This includes not only null findings or unexpected findings but also studies that are considered to have failed, with careful explanation of the circumstances of the failure (e.g., experimental error, failed manipulation check). The context in which these data have been collected and whether the data are connected to published studies (e.g., dropped experiments) are carefully explained.

Making a set of items appear to contain more or fewer items than it actually has simply by manipulating their variety (e.g., whether the items have different shapes or colors) can be of use in many different settings. For example, the serving size of a meal can be made to appear large for a customer in a restaurant or the number of people in a demonstration may be perceived as large when there is variety. Whether the fries in a serving all have the same shape or whether the people in a demonstration wear the same shirts can easily be varied to create the desired effect. On the basis of the principles of Gestalt psychology, Redden and Hoch (2009) argued that quantities, such as the number of objects, appear smaller in number if they are heterogeneous (i.e., varied in terms of colors or shapes) than if they are homogeneous (e.g., single-colored). We call this the *variety effect*. The argument is that perceiving a set of items as a whole makes it appear larger in number. In four studies, Redden and Hoch (2009) asked people to estimate the numbers of objects in matrices (Studies 1 and 2) and to pour M&M's (Studies 3 and 4) and found that variety in the items reduced the perceived quantity with a very large average effect size of d = 1.105, 95% CI [0.886, 1.324]. Variety was manipulated by including one or multiple colors or shapes. The effect did not depend on the "strength" of the variety, that is, whether there were two colors or shapes or more than two. In line with the Gestalt argument, arranging dots in a pattern (vs. randomly), as would be the case in a Gestalt, has led people to estimate sets as larger in number (e.g., Ginsburg, 1978, see also Frith and Frith, 1972; Vos et al., 1988).

However, findings on the relationship between variety and perceived quantity have not been consistent in recent research. In line with Redden and Hoch (2009) findings, Dakin et al. (2011, p. 19554, Figure 3J) found that a set of black and white dots appeared smaller in number than a set of white dots. However, they did not provide data or effect sizes for the difference. Zhao and Yu (2016) found the opposite: They organized (vs. randomized) how different items appeared in a set and thus facilitated the perception of a whole Gestalt. Across four studies, they did not find that heterogeneous sets of items were perceived as smaller in number than homogeneous sets of items. More specifically, Zhao and Yu (2016) Experiment 1a revealed a very small or imprecise effect (p = .03) of heterogeneity, and their Exper-

iments 1b, 2, and 3 revealed no variety effect. Experiments 1a, 1b, and 2 revealed an interaction with variety and actual quantity. Accordingly, varied sets of items were perceived as larger in number if they were large in number (i.e., 20 items) but not if they were small in number (i.e., 10 items). Finally, Burr et al. (2017) found that the estimates of items with very high density resulted in a set being perceived as a texture. In their study, high density was associated with estimates of a large number of items but a severe underestimation of the number of items. By contrast, if there were different items (vs. identical items) in a set, the set was not perceived as a texture as easily. Based on this argument and contrary to the prediction made by Redden and Hoch (2009), variety may *increase* the perceived quantity for large numbers as it prevents the objects from "blurring" into a single texture.

Across three consecutive years (2012 to 2014), we conducted six replications of the studies reported by Redden and Hoch (2009). We report our studies in chronological order. The studies were part of student projects that were supervised by a professor. The studies' research questions went beyond the variety effect to some extent. In this report, however, we tested the variety effect in all results sections first. All available study materials and data sets can be found online (https://osf.io/9s4w7). We invite other researchers to reanalyze our data or to conduct studies using our materials and to replicate the results. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (Simmons et al., 2012). In all studies, we used R version 4.2.1 (R Core Team, 2018) to analyze the data and R functions provided by Wolf (2021) to report the results. The analysis code for all studies is available online (https://osf.io/jtn4f).

Study 1

To find out whether variety reduces perceived quantity, we asked participants to estimate the quantities of single-colored or multicolored Nestlé Smarties (this is the European equivalent to M&M's, different from the Smarties found in the U.S.) in a between-subjects design. Note that Redden and Hoch (2009) chose to manipulate variety between-subjects in the pouring studies (Studies 3 and 4) and within-subjects in the quantity estimation studies (Studies 1 and 2). We chose a between-subjects manipulation for our studies that featured Smarties because we wanted to prevent demand characteristics: The well-known advertisement slogan of Smarties is "many many colorful Smarties" (in German "viele viele bunte Smarties"). Thereby, the German word "bunt" explicitly refers to something having more than one color (e.g., varied; colorful can also be understood as an antonym to colorless instead of single-colored). We were concerned participants would be made aware of that slogan and accordingly report that colorful Smarties are more numerous.

Besides the photographs of 144 multicolored Smarties in each of the four containers, we created photographs of only brown, blue, green, purple, pink, red, orange, and yellow (eight single-color conditions). We went beyond the original study as we manipulated the container in which the Smarties were presented (small round bowl vs. rectangular bowl vs. large round bowl vs. slim jar) in a within-subjects design. Container sizes were previously identified as another predictor of estimate biases (Kahn & Wansink, 2004), and we were curious about how this would interact with the variety effect. On the basis of findings by Wansink and van Ittersum (2003), we hypothesized that the height of the containers would lead to an overestimation of the quantity. That is, estimates should be largest for the slim and high container (Number 3 in Figure 1). Another deviation from the original studies was the use of an anchor (e.g., Tversky and Kahneman, 1974): Before asking participants how many Smarties were in the different bowls (free-text entry), we asked them "was the number of Smarties more than or less than 100?" They had to check either "more" or "less" We expected that such an anchor would reduce the variance of the estimates.

Figure 1

Containers used in Studies 1 and 2 with Homogenous Sets of Smarties

Figure 2

Containers used in Studies 1 and 2 with Multicolored Sets of Smarties

Participants were randomly assigned to one of the color groups (half were presented the multicolored sets and half were presented one of the eight single-colored sets of Smarties). The photographs of the Smarties in the four containers were presented in a random order. After viewing each photograph, participants (a) estimated the number of Smarties in each container, (b) stated how much fun they would have eating that many Smarties, and (c) stated whether they would like to eat from the container. The last two questions were asked because we expected participants to prefer consuming a heterogeneous set of Smarties (Kahn & Wansink, 2004) and ultimately wanted to unconfound whether consuming a larger quantity of multicolored Smarties would be due to their appearing to be less numerous or due to their assortment appearing more pleasant. At the end of the study, participants provided demographic data (sex, age, and how often they actually consumed Smarties).

Method

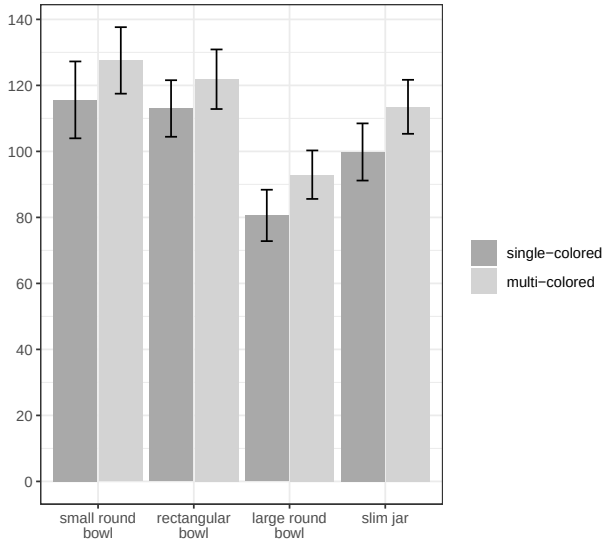
We conducted the study online in March 2012 over the course of 3 days and recruited mainly students. There were N = 276 participants. We excluded all participants who had missing values for at least one of the quantity estimates and thus ended up with a final sample of 206 participants (118 of which were female and 102 of which saw the single-colored candy). Their mean age was M = 23.26 years (SD = 7.39 years). The questionnaire was programmed with Equip Questionnaire Generator (Lutsch, 2001). No power analysis was conducted beforehand. As this study was part of a student project and data had to be collected in a short amount of time, recruitment was limited to 3 days, and we recruited as many participants as possible. As this study was a conceptual replication, we were able to compute the achieved power given the overall effect size of all four studies reported by Redden and Hoch (2009), that is, d = 1.105 (N_{total} = 306). A post hoc power analysis indicated that the power was sufficient (1 - β > .999). Power analyses for all studies were computed with G*Power (Faul et al., 2007). Power protocols for all studies are available online (https://osf.io/v6p28/).

Result

For our main analysis, we averaged the estimated quantity of Smarties across all bowls. Internal consistency across bowls was high (Cronbach's α = .866, four items). Overall, single-colored quantities were estimated to be smaller in number (M = 102.26, SD = 40.09, N = 102) than multicolored quantities (M = 113.97, SD = 37.43, N = 104), t(204) = -2.17, p = .031 (two-tailed), d = -0.302. Note that the hypothesis

Figure 3

Mean Estimates of 144 Smarties by Variety and Container



Note. Error bars represent 95% confidence intervals.

was actually one-tailed (variety reduces perceived quantity). A repeated-measures ANOVA with a 2 (variety: single-colored vs. multicolored; between-subjects) × 4 (container: small round bowl vs. rectangular bowl vs. large round bowl vs. slim jar; within-subjects) further revealed an effect of container, F(3, 612) = 60.63, p < .001, η^2 = .229 (see Figure 3). For example, the estimated quantity in the rectangular bowl was significantly larger than the estimated quantity in the large round bowl. There was no interaction between container and variety, F(1, 612) = 0.26, p = .854, η^2 = .001. To account for the violation of the normality assumption in the data set (i.e., significant Shapiro-Wilk normality test, W = 0.95, p < .001) and them being count data, we additionally conducted a Wilcoxon rank sum test with continuity correction for the relationship between quantity estimates and variety, which converged with the results from the t test, W = 4173.5, p = .008 (two-tailed).

Discussion

In this study, participants estimated the number of Smarties in glass containers with different shapes in either a multicolored condition or one of eight single-colored conditions. Contrary to our hypothesis, variety *increased* instead of decreased the perceived quantity. Note that this is a conceptual replication. Our study differed from the studies by Redden and Hoch (2009) in that (a) we used Smarties instead of M&M's, (b) par-

ticipants estimated the number of pieces of candy from looking at a photograph rather than pouring them, (c) participants were presented with an anchor before making each estimate, (d) participants did not receive remuneration or course credit, and (e) the study was conducted online, not in a laboratory.

Study 2

Study 2 was a close replication of our Study 1. It was conducted to estimate the robustness of the empirical effect size that was in the opposite direction of the hypothesized effect. In addition to the multicolored condition and the single-colored conditions, there were six more conditions in which people saw 144 smarties that had two colors. These color combinations were blue-brown, blue-orange, blue-red, green-brown, green-purple, and green-red. Half of the participants saw one of the eight single-colored conditions and half of them saw one of the six bicolored conditions or the multicolored condition. Everything else remained unchanged.

Method

During the recruitment conducted in the lab on June 9, 2012, which was the university's open day and attracted young people interested in studying but also their parents and alumni. A total of 174 participants completed the study. After we advertised the online version of the study among students. There were N = 610 people who clicked on the link to the study and N = 439 complete cases (255 female, mean age 24.22 years). After one participant made an estimate of over 4 million Smarties (144 was the true number of Smarties), we decided to exclude estimates that exceeded the mean estimate by ±4 SD as had been done in Study 3 by Redden and Hoch (2009, p. 412). The questionnaire was programmed with Equip Questionnaire Generator (Lutsch, 2001). Again, no power analysis was conducted beforehand. Recruitment was limited to 1 day in the lab but was then continued using an online version of the study. The power to conceptually replicate the effect size of all four studies reported by Redden and Hoch (2009), that is, d = 1.105 (N_{total} = 306), was very high (1 - β > .999). The power to replicate the effect from Study 1, that is, d = -0.302, was sufficient, too (1 - β > .933).

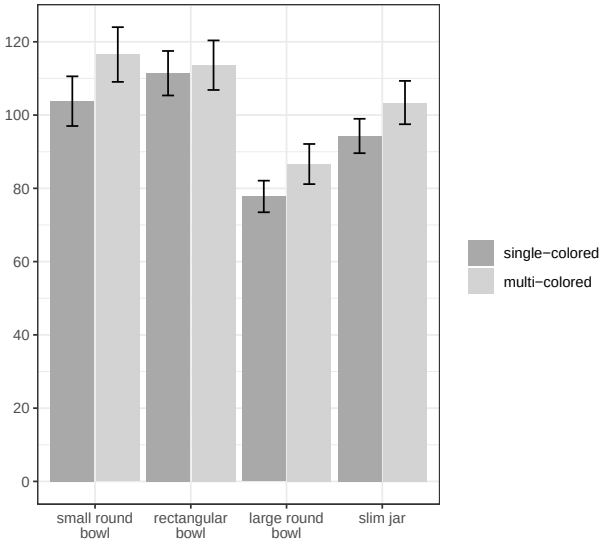
Result

To test the variety hypothesis, we averaged the estimated quantity of Smarties across all bowls. Internal consistency was high (Cronbach's α = .875, four items). Overall, the single-colored quantities were estimated to be smaller in number (M = 96.85, SD = 36.79, N = 240) than the multicolored quantities (M = 105.26, SD = 40.57, N = 199), t(437) = -2.28, p = .023 (two-tailed), d = -0.218. A repeated-measures ANOVA with a 3 (variety: single-colored vs. bicolored vs. multicolored; between-subjects) × 4 (container: small round bowl vs. rectangular bowl vs. large round bowl vs. slim jar; within-subjects) corroborated the variety effect, F(1, 435) = 4.31, p = .038, η^2 = .010, and further revealed an effect of container, F(3, 1305) = 118.16, p < .001, η^2 = .214 (see Figure 4) that was similar to the effect in Study 1. For example, the quantities were estimated to be larger in the rectangular bowl than in the large round bowl. There was no interaction between container and variety, F(3, 1305) = 2.42, p = .065, η^2 = .006.

To account for the violation of the normality assumption in the data set (i.e., significant Shapiro-Wilk normality test, W = 0.93, p < .001) and them being count data, we additionally conducted a Wilcoxon rank sum test with continuity correction for the relationship between quantity estimates and variety, which converged with the results from the t test, W = 20,856, p = .022 (two-tailed).

Figure 4

Mean Estimates of 144 Smarties by Variety and Container



Note. Error bars represent 95% confidence intervals.

Discussion

Study 2 successfully replicated the effect from Study 1 because variety resulted in an increase rather than a decrease in perceived quantity, although the effect might have been driven by the fact that bi-colored quan-

tities descriptively appeared larger in number than both single- and multicolored quantities. Moreover, the effects of the containers on the estimated quantities were successfully replicated.

Study 3

The results of the previous studies were puzzling with regard to why the effect we found was different from the hypothesized effects. In Study 3, we sought to stick even more closely to the procedure used by Redden and Hoch (2009) to replicate the original findings. To also explain our opposing findings, we varied the quantity of the presented set of Smarties to appear either two-dimensional or three-dimensional because this was one of the differences between the quantities presented by Redden and Hoch and our approach: The Smarties we presented were in three-dimensional containers in most cases, whereas Redden and Hoch used two-dimensional matrices in their Studies 1 and 2. In Study 3, they used 6-inch-diameter bowls with 52 grams of M&M's (p. 412) so that the quantity was most likely[1] presented on a plane and in Study 4, they used 9-inch-diameter plates (p. 413). Three-dimensional single-colored sets of things might appear two-dimensional due to blurring borders between objects, whereas three-dimensional multicolored sets would allow the viewer to infer the depth due to the different colors. Recently, Burr et al. (2017) have argued similarly, that textures in sets of stimuli with high density might prevent the stimuli's quantity from being underestimated.

Method

We had participants look at the Smarties on a plate and then asked them to pour the same number of Smarties onto an empty plate. In doing so, we manipulated the variety of the Smarties we presented (one vs. four colors), the quantity and thereby the arrangement of the Smarties (32 flat or two-dimensional vs. 116 in a pile or three-dimensional), and the variety of the Smarties that were poured (one vs. four colors) in an incomplete within-subjects design. Note that arrangement and quantity could not be manipulated independently without changing the container, which also mattered. Thus, arrangement was confounded with quantity. However, there is no reason why quantity would moderate the variety effect. Smarties were poured from a transparent bottle that included a total of 248 pieces. That is, every participant completed four trials, which included a small single-colored, a small multicolored, a large single-colored, and a large multicolored set of Smarties. However, not every set was estimated by pouring both single-colored and multicolored Smarties. Instead, participants were randomly assigned to one of two groups, and this determined the two sets for which they had to use the bottles with single-colored Smarties and for which they had to use the bottles with multicolored ones. The order of these sets was completely randomized.

Participants estimated the number of Smarties that had been presented by pouring from the single- or multicolored bottle of Smarties and by estimating the number afterwards. The candy that was poured was weighed and written down by the experimenters after the participants had completed the experiment. Before estimating the sets of Smarties and pouring their estimates, participants were told what to do and completed a test run to see how to handle the bottle from which they were had to pour the Smarties. Prior to each estimate, we asked participants if there were more or less than 80 Smarties (anchor). Figure 5 provides an overview of the study procedure. Before estimating the number of Smarties, participants filled out a small form with information on their sex, age, occupation, and whether they were familiar with the advertising slogan "many many colorful Smarties."

Participants were tested in the University lab on June 8, 2013, which was the university's open day and attracted young people interested in studying but also their parents and alumni. There were 144 participants, 89 of whom were female (mean age = 29.6 years). There were no exclusions.

Again, no power analysis was computed beforehand. Recruitment was limited to 1 day. The power to conceptually replicate the effect size for the two candy-pouring studies reported by Redden and Hoch (2009), that is, d = 0.489 (N_{total} = 169), was very high (1 - β > .999).

Results

The weight of the Smarties that were poured had low internal consistency (Cronbach's α = .560, four items), and the estimates had acceptable internal consistency (Cronbach's α = .702, four items).

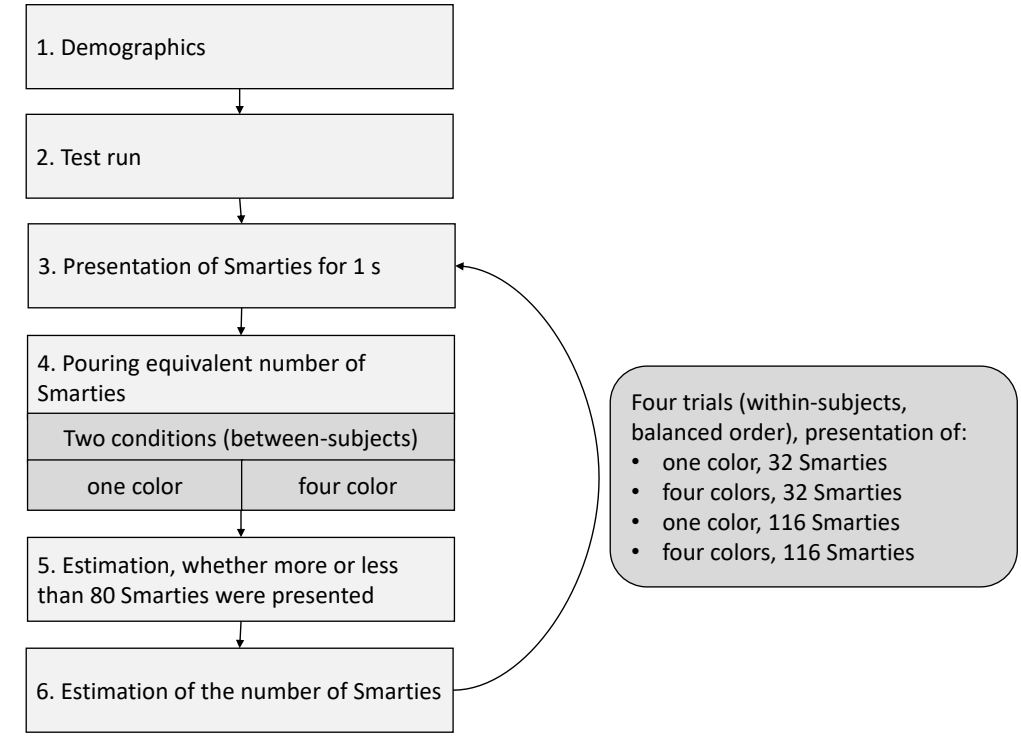
We tested the hypothesis separately for the two dependent variables (weight and estimated quantity), but this resulted in an inflation of the alpha level, and thus, the criterion for significance that we chose to use was α < .025 (Bonferroni-corrected).

If variety reduces perceived quantity, participants should pour less Smarties if they are single-colored (because they look like more). However, the weight of the poured Smarties was not significantly lower for single-colored sets of Smarties (M = 109.09 g, SD = 16.11 g,

[1] Prior to the experiment, we tried to pour the Smarties into a pile and found that as long as there was space at the bottom of a bowl, the Smarties never sat on top of each other.

Figure 5

Procedure of Study 3



N = 144) than for multicolored ones (M = 107.20 g, SD = 17.20 g), t(143) = 1.37, p = .172 (two-tailed), dz = 0.114. The estimated number of Smarties was not significantly larger for the single-colored sets of Smarties (M = 70.39, SD = 34.99) than for the multicolored ones (M = 66.84, SD = 30.28), either, t(143) = 1.82, p = .070 (two-tailed), dz = 0.152. The estimated number of poured Smarties was not influenced by the variety of the colors of the Smarties (M = 109.73, SD = 58.67 vs. M = 106.56, SD = 52.78), t(143) = 0.35, p = .725 (two-tailed), dz = 0.029.

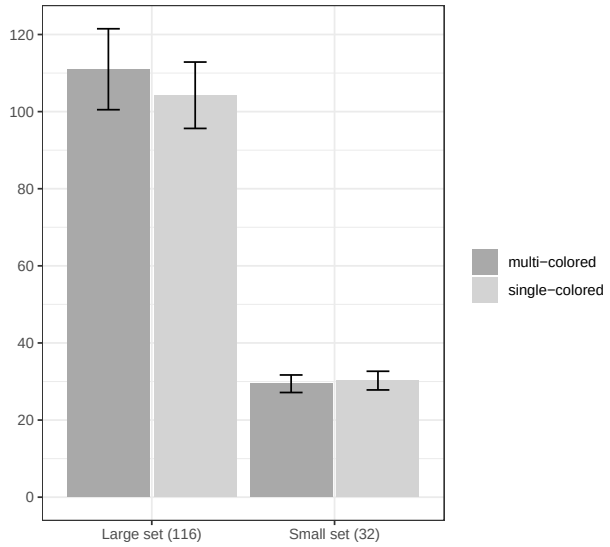
To test the dimensionality hypothesis, we used the number estimates because they had higher internal consistency, and we conducted a 2 (variety: single-colored vs. multicolored; within-subjects) × 2 (quantity: 32 vs. 116; within-subjects) × 2 (pouring color: single-colored vs. multicolored; between-subjects) mixed ANOVA. Whereas the quantity of the Smarties had a large effect, F(1, 141) = 365.54, p < .001, η^2 = .722, there was no main effect of variety, F(1, 141) = 2.98, p = .087, η^2 = .021. The only other effect of the model that reached statistical significance was a small interaction between variety and quantity, F(1, 141) = 5.53, p = .020, η^2 = .038 (see Figure 6). Post hoc tests revealed that there was no effect of variety in the small set, t(143) = 0.64, p = .521, dz = 0.05, 95% CI [-0.11, 0.22], and there was a small effect in the large set, t(142) = -2.14, p = .034, dz = -0.18, 95% CI [-0.34, -0.01]. To account for the quantity estimates' violation of the normality assumption (i.e., significant Shapiro-Wilk normality test, $W_{variety}$ = 0.90, $W_{no\ variety}$ = 0.86, both ps < .001) and them being count data, we additionally conducted a Wilcoxon signed rank test with continuity correction for the relationship between quantity estimates and variety, which converged with the results from the t test, V = 5,093, p = .205 (two-tailed).

Discussion

We tried to consolidate the findings from our Studies 1 and 2 and the opposite effects reported by Redden and Hoch (2009) by (a) using methods that closely mirrored the original experiments and (b) including manipulations such as the arrangement of the Smarties on a plane versus in a pile (or a small vs. a large set) to determine whether this was the reason for the opposite effects. In the large set, we found a very small effect that variety *increased* instead of *decreased* perceived quantity and in the small set we found no effect of variety.

Figure 6

Mean Estimates of Smarties by the Variety and Quantity of the Set



Note. Error bars represent 95% confidence intervals.

Study 4

After conceptual replications produced the opposite effect (Studies 1 and 2), and after the replication of Redden and Hoch (2009) pouring studies failed to produce the hypothesized effect (Study 3), we chose to replicate Redden and Hoch's Study 1. Thus, participants had to estimate the number of symbols in a matrix after we manipulated the variety of the symbols or the symbols' colors. For exploratory purposes, we also varied the presentation time. Note that this time, there was no anchor[2].

Method

We created 50 matrices consisting of 100 boxes that were presented for 750 or 1,500 ms. The matrices were varied with respect to how many boxes were filled with a shape or color (from 30% to 70% in steps of 10%), what was in the boxes (colors vs. shapes), and how many different colors or shapes were in the boxes (colors: red, green, blue, orange, and purple; shapes: triangle, rhombus, star, square, and smiley). Note that for each combination (e.g., 30% red and green boxes), there was only one picture. In Redden and Hoch's study, the stimuli were created anew for each trial, whereas we had 50 fixed stimuli that were used for all participants. Afterwards, we asked participants how many boxes were filled. In contrast to our other studies, there was no anchor. Examples of four matrices are presented in Figure 7. The matrices were created in Excel (see https://osf.io/b57gt/ for the generation program and all the stimuli we used). The questionnaire was programmed with Inquisit Millisecond version 3 ("Inquisit 3," n.d.).

All factors in the design were manipulated within subjects. Matrices were presented in a randomized order. Due to a programming error, the 50 matrices were not drawn without replacement so that each participant saw each matrix but with replacement. In cases in which a matrix was drawn multiple times, only the last estimate was saved, leading to missing values for approximately 18 stimuli per participant.

Participants were recruited online between November 26, 2013 and January 15, 2014 via social networks and the university newsletter. There were 98 people who clicked on the link to the study and 82 complete cases (52 female, mean age = 27 years). Again, no power analysis was computed beforehand. Recruitment was limited by the course deadline. The power to directly replicate the effect size for the quantity-estimation studies reported by Redden and Hoch (2009), that is, d = 1.866 (N_{total} = 137), was very high (1 - β > .999).

Results

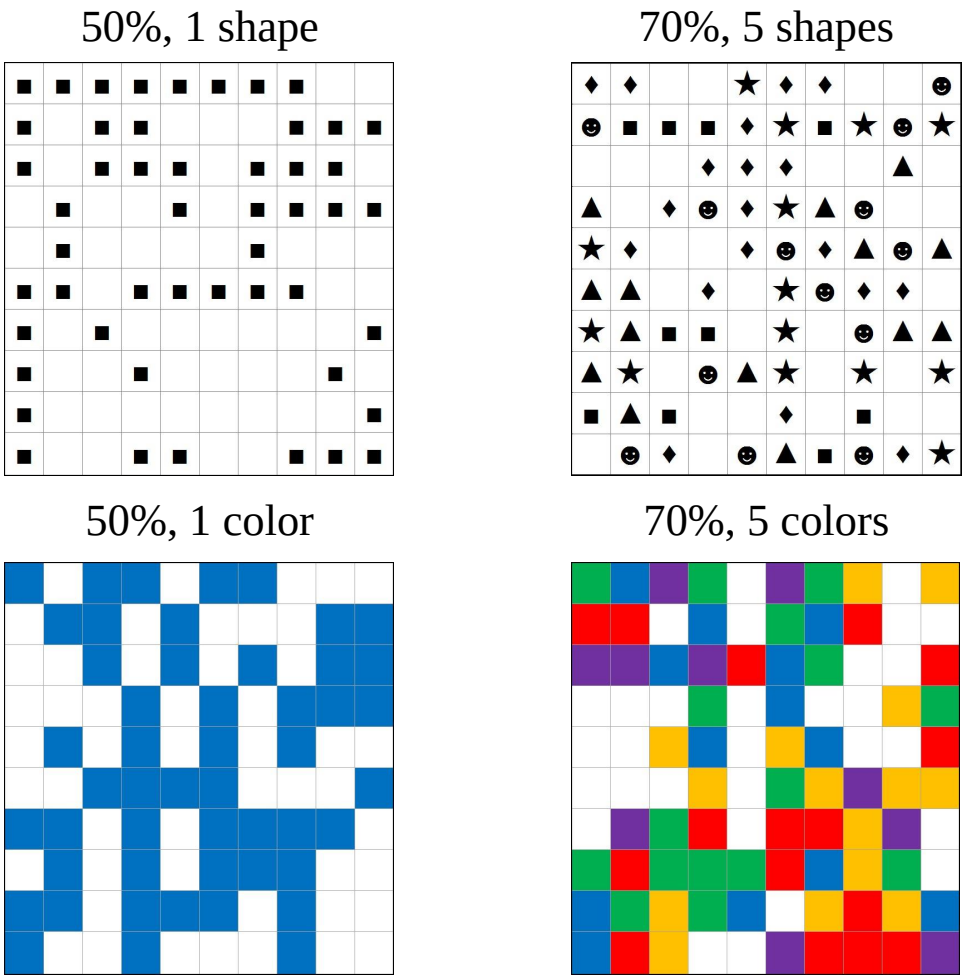
Due to the programming error, we could not run our planned analysis (i.e., a mixed-effects ANOVA with a quantity × variety × type of symbol × presentation time design). Instead, we aggregated all estimates of homogenous matrices and all estimates of heterogeneous matrices, respectively. They were strongly correlated, r(80) = .711, p < .001. Homogenous matrices (M = 54.31, SD = 9.81, N = 82) were estimated to contain slightly fewer elements than heterogeneous matrices (M = 54.58, SD = 8.29, N = 82), although this difference was not significant, t(80) = -0.35, p = .727 (two-tailed), dz = -0.039. Exploratory analyses revealed no effect of presentation time.

To account for the violation of the normality assumption in the data set (i.e., significant Shapiro-Wilk normality test, $W_{varietyl}$ = 0.88, $W_{varietyl}$ = 0.94, both ps < .001) and them being count data, we additionally conducted a Wilcoxon signed rank test with continuity correction for the relationship between quantity estimates

[2]In a personal correspondence with Joseph P. Redden, he explained that he deliberately avoided the use of anchors because he suspected that the subtle difference of choosing an anchor might overshadow or reverse the effect because people give more weight to the number of types cues instead of the Gestalt area.

Figure 7

Examples of the Matrices used in Study 4



Note. See https://osf.io/75bny/ for all stimuli.

and variety, which converged with the results from the t test, V = 1,662, p = .996 (two-tailed).

Discussion

In an attempt to conduct a close replication of Study 1 by Redden and Hoch (2009), we failed to replicate the effect of variety on perceived quantity. Although there was a programming error, which led to an alternative analysis of the results, power was not compromised. At this point, deviations from the original study consisted of the participants' nationality (from the U.S. vs. from Germany) and the specific stimuli. Neither factor was found to explain the effect, however, namely that homogenous sets of entities form a Gestalt, which in turn seems larger.

Study 5

This study was a direct replication of our Study 4, except that the programming error had been fixed, and the presentation time was set to 750 ms for all trials.

Method

Participants were recruited online from social networks and the university newsletter in September 2014. There were 50 participants and 45 complete cases (38 female, mean age = 21.34 years). Again, no power analysis was computed beforehand. Recruitment was limited by the course deadline. The power to directly replicate the effect size for the quantity-estimation studies reported by Redden and Hoch (2009), that is, d = 1.866 (N_{total} = 137), was very high (1 - β > .999).

Results

As in Study 4, we aggregated all estimates of homogenous matrices and all estimates of heterogeneous matrices, which were strongly correlated, r(43) = .887, p < .001. The homogenous matrices (M = 52.62, SD = 13.60, N = 45) were estimated to contain slightly fewer elements than the heterogeneous matrices (M = 52.09, SD = 11.68, N = 45), although this difference was not significant, t(44) = 0.57, p = .574 (two-tailed), dz = 0.084.

Furthermore, we computed a 5 (percentage filled) × 5 (number of different objects) × 2 (kind of object: shape vs. color) repeated-measures ANOVA. There was a large effect of percentage filled on the perceived quantity, F(1.44, 44.50) = 107.98, p < .001 (Greenhouse-Geisser-corrected due to sphericity), η^2 = .777. The number of different objects, that is, the variety, had no effect on the perceived quantity, F(3.05, 94.42) = 0.95, p = .419 (Greenhouse-Geisser-corrected due to sphericity), η^2 = .030. The analysis further revealed an unpredicted percentage filled × number of different objects interaction, F(9.46, 293.20) = 3.10, p = .001 (Greenhouse-Geisser-corrected due to sphericity), η^2 = .091. The conditions involved in the complex interaction are presented in Figure 8.

To account for the violation of the normality assumption in the data set (i.e., significant Shapiro-Wilk normality test, $W_{varietyl}$ = 0.82, $W_{varietyl}$ = 0.87, both ps < .001), we additionally conducted a Wilcoxon signed rank test with continuity correction for the relationship between quantity estimates and variety, which converged with the results from the t test, V = 530, p = .495 (two-tailed, two ties).

Discussion

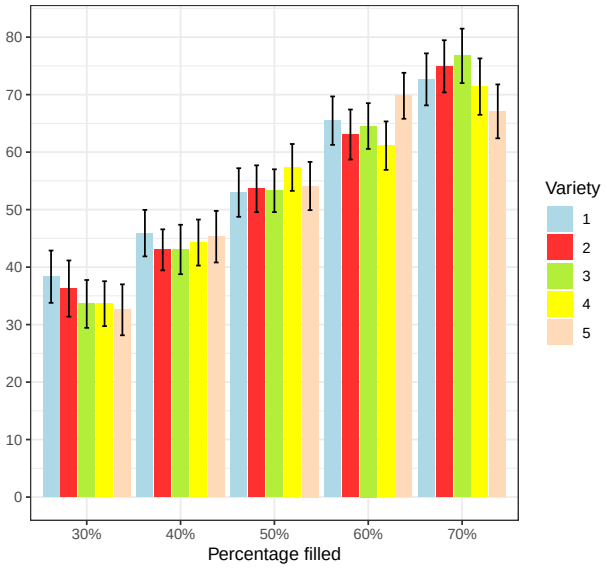
In another attempt to conduct a close replication of Study 1 by Redden and Hoch (2009), we again failed to replicate the effect of variety on perceived quantity.

Study 6

In this study, we again replicated our Study 1 but dropped the container factor by using only one large bowl. Instead of using the photographs of the different single- and multicolored Smarties from the previous studies, we manipulated the photograph of the multicolored Smarties to appear single-colored (see Figure 9). Thus, the arrangement of the candy in the bowl was held constant for all colors and assortment and variety were not confounded. In the bowl, the Smarties appeared as a flat surface.

Figure 8

Mean Estimates of Boxes in Matrices by Variety (Number of Colors or Shapes) and Quantity (Percentage of Filled Boxes) of the Set



Note. Error bars represent 95% confidence intervals.

Method

Recruitment was conducted online via social networks and the university newsletter from August to November 2014 and yielded 259 participants. Participants were excluded if they had missing values for at least one of the estimates, or gave estimates of 1 or 0 for the number of Smarties in the bowl. A total of 161 participants did not meet the exclusion criteria. The questionnaire was programmed with Equip Questionnaire Generator (Lutsch, 2001).

Again, no power analysis was computed beforehand. As this study was a conceptual replication, we were able to compute the achieved power given the overall effect size for all four studies reported by Redden and Hoch (2009), that is, d = 1.105 (N_{total} = 306). A post hoc power analysis indicated that the power to replicate the total effect was sufficient (1 - β > .999). The power to replicate the effect from our Study 1, that is, d = -0.302, was low (1 – β = .603).

Results

Overall, the single-colored quantities received lower estimates (M = 79.14, SD = 45.14, N = 77) than the multicolored quantities (M = 93.81, SD = 52.18, N = 84), t(159) = -1.90, p = .059 (two-tailed), d = -0.301. To account for the violation of the normality as-

Figure 9

The Eight Colors used in Study 6 and the Multicolored Set of Smarties

sumption in the data set (i.e., significant Shapiro-Wilk normality test, W = 0.86, p < .001) and them being count data, we additionally conducted a Wilcoxon rank sum test with continuity correction for the relationship between quantity estimates and variety, which had a slightly lower (and thus significant) p-value than the t test, W = 2,619, p = .037 (two-tailed).

Discussion

In this study, we again failed to replicate the hypothesized effect that variety would reduce perceived quantity. Furthermore, the opposite effect reported in Study 1 of this research article was not clearly replicated either, which might be because power was too low ($1 - \beta$ = .603) to detect it.

Table 1

Overview of original and replication study features

Study	IV	DV	Manipulation of variety	Dimension of variety	Quantities	Variety levels	Stimulus type	Stimulus presentation time [ms]	Anchor
RH 1	Variety	Estimation of symbols (#/%)	Within-subjects	Color, shape	30, 40, 50, 60, 70	2, 3, 4, 5	Unfilled symbols in a matrix	750	-
RH 2	Variety	Estimation of symbols (#/%)	Within-subjects	Color, shape	30, 40, 50, 60, 70	2, 3, 4, 5	Unfilled symbols in a matrix	750	-
RH 3	Variety of poured candy	Pouring candy	Between-subjects	Color	52 g	3	M&M's	Unlimited	-
RH 4	Variety of poured and seen candy	Pouring candy	Between-subjects	Color	55 g, 66 g	4	M&M's	Unlimited	-
1	Variety	Estimation of candy (#)	Between-subjects	Color	144	8	Photographs of Smarties	Unlimited	100
2	Variety	Estimation of candy (#)	Between-subjects	Color	144	2, 8	Photographs of Smarties	Unlimited	100
3	Variety of seen or poured candy	Estimation of candy, Pouring candy (#)	Within-subjects	Color	32, 116	4	Real Smarties	1000	80
4	Variety	Estimation of symbols (#/%)	Within-subjects	Color, shape	30, 40, 50, 60, 70	5	Filled symbols in a matrix	750 or 1500	-
5	Variety	Estimation of symbols (#/%)	Within-subjects	Color, shape	30, 40, 50, 60, 70	5	Filled symbols in a matrix	750	-
6	Variety	Estimation of candy (#)	Between-subjects	Color	144	8	Photographs of Smarties	Unlimited	100

Note. RH: Redden and Hoch, IV: Independent variable, DV: Dependent variable, #: Estimates were given as a number, #/%: Estimates were given as a number between 0 and 100, which was equivalent with the percentage of filled boxes in the matrices.

General Discussion

In six studies that included three close and three conceptual high-powered replications, we could not replicate the effect that variety reduces perceived quantity. An overview of the differences between the original and the replication studies is provided in Table 1. The results of the original studies and our replications are summarized in Table 2 and Table 3, respectively.

Meta-analysis of Effect Sizes

The meta-analytical effect that Redden and Hoch (2009) found was positive and significant, $t(4) = 2.96$, $p = .042$, $d = 1.12$, 95% CI [0.070, 2.169], $N_{total} = 306$. There was no effect in our replication studies, $t(7) = -1.16$, $p = .284$, $d_{replication} = -0.075$, 95% CI [-0.227, 0.077], $N_{total} = 1{,}077$. The difference between the effects from the original studies versus the replication effects was significant, $F(1, 11) = 17.78$, $p = .001$. The overall effect of all 10 studies and 13 reported findings was not significantly different from zero, $t(12) = 1.49$, $p = .161$, $d = 0.355$, 95% CI [-0.163, 0.873], $N = 1{,}383$ (see Figure 10 for a forest plot). Meta-analyses were conducted using R (R Core Team, 2018) and the packages psych (Revelle, 2018), readxl (Wickham & Bryan, 2018), forestplot (Lumley & Gordon, 2019), dmetar (Harrer et al., 2019), and meta (Balduzzi et al., 2019).

Note that the parametric effect size estimates synthesized here violate the normality assumption in most studies and a synthesis of non-parametric estimates would be more appropriate. However, we chose to use Cohen's ds to allow for comparability between studies. Moreover, in our data, results from both estimates strongly converged.

Presence of Questionable Research Practices in our Replications

All of the original studies were conducted during a time in which the presence of questionable research practices had either been an integral part of psychological science or had just begun to be noticed. Most researchers had not yet realized that most published research could not be replicated (cf. Ioannidis, 2005). Although it is unclear whether there was any motivation to engage in p-hacking (e.g., Simmons et al., 2011) in any of the studies, the possibility that this might have occurred should be discussed.

Note that it is possible to engage in p-hacking in multiple ways. Not only can researchers try out different methods to see if they can obtain a significant result in the opposite direction of the original hypothesis, but they can also try to get a null result by engaging in so-called null-hacking (Yeager et al., 2019). P-curve analyses to test for p-hacking do not work for nonsignificant results (i.e., they do not work to determine null-hacking, either). We want to reject potential accusations of null-hacking on the grounds that we tried very hard to get the effect over many years and never intended to fail to find the effect. If we wanted to show that there was no effect, why would we have tested numerous moderators and conducted further studies? Also, at the time, null results were extremely difficult to publish. Neither the original studies nor our replication attempts were preregistered or included a power analysis. As is the case for almost all studies from that time, this does not meet the current standards for experimental research. We attempted to solve these problems by reporting all studies and all measures (including a study that had a minor programming error) by sharing our data and analysis code (https://osf.io/9s4w7) and by computing post hoc power analyses, which yielded high power in all cases given the original effect sizes. One researcher degree of freedom (e.g., Wicherts et al., 2016) that could not be ruled out by this degree of transparency was the addition of further observations. There were instances in our studies when planned periods of participant recruitment were extended (Study 2). However, excluding Study 2 from the meta-analysis did not change the finding that the total effect was 0.

Future Direction

In six different studies, we failed to replicate the effect that variety reduces perceived quantity. Based on our results, we deem moderators, such as the "dimensionality" of the sets (2D planes vs. 3D piles), the presentation time (750 ms vs. 1,000 ms vs. 1,500 ms), and the kind of item that was estimated (symbols vs. colors vs. Smarties) unlikely. Another possible confound was participants' nationality (American vs. German). Moreover, our studies differed in that the code that was used to produce the stimuli differed from the original studies, and we used Smarties instead of M&M's. To allow for closer replication, we recommend future research to provide or reuse stimuli or code for stimuli generation. Less noisy measurements might also be achieved by using two-alternative-forced-choice response formats instead of letting participants estimate the number of items. Furthermore, we recommend future researchers to conduct more complex but more appropriate analyses. Using multi-level modelling, design factors such as the container or color can adequately be represented in the statistical models.

However, none of these differences has a meaningful theoretical connection to the reasoning behind the effect, so either these aspects do not impact whether the effect occurs or the theoretical account is wrong.

Table 2

Effect Sizes from the Original Studies Conducted by Redden and Hoch (2009)

Study	Page	IV	DV	Reported effect	d	N
RH 1	409	Variety	Estimation of symbols	F(1, 79) = 110.19, p < .0001 (M = 56.8 vs. 53.5)	2.377	80
RH 2	411	Variety	Estimation of symbols	F(1, 56) = 18.13, p < .0001	1.148	57
RH 3	412	Variety of poured candy	Pouring candy	F(1, 99) = 4.00, p < .05 (M = 59.2 vs. 53.0)	0.394	105
RH 4	414	Variety of poured candy	Pouring candy	F(1, 63) = 6.45, p < .02	0.645	64
RH 4	414	Variety of seen candy	Pouring candy	F(1, 63) = 6.48, p < .02	0.646	64

Note. Positive effect sizes indicate that variety reduced perceived quantity.

Table 3

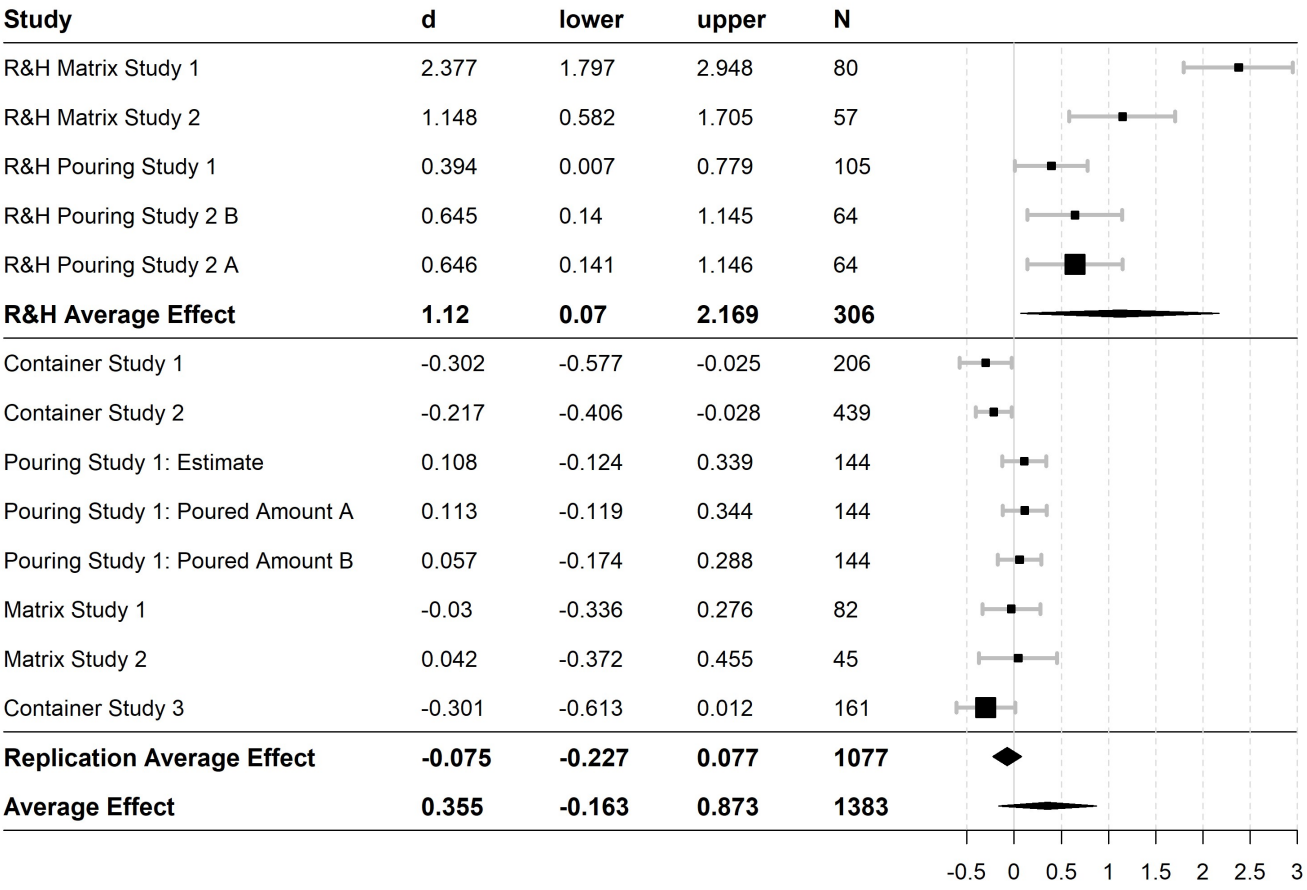
Effect Sizes from the Replication Studies

Study Design	IV	DV	Reported effect	d	$N_{variety}$	$N_{no\ variety}$
1	Variety Estimation of candy	t(204) = -2.17, p = .031 (two-tailed), d = -0.302	-0.302	102	104	Between-subjects
2	Variety Estimation of candy	t(437) = -2.28, p = .023 (two-tailed), d = -0.217	-0.217	240	199	Between-subjects
3	Variety of seen candy Estimation of candy	t(143) = 1.82, p = .070 (two-tailed), dz = 0.152	0.108	144	144	Within-subjects
3	Variety of seen candy Pouring candy	t(143) = 1.37, p = .172 (two-tailed), dz = 0.114	0.113	144	144	Within-subjects
3	Variety of poured candy Pouring candy	t(143) = 0.35, p = .725 (two-tailed), dz = 0.029	0.057	144	144	Within-subjects
4	Variety Estimation of symbols	t(80) = -0.35, p = .727 (two-tailed), dz = -0.039	-0.030	82	82	Within-subjects
5	Variety Estimation of symbols	t(44) = 0.57, p = .574 (two-tailed), dz = 0.084	0.042	45	45	Within-subjects
6	Variety Estimation of candy	t(159) = -1.90, p = .059 (two-tailed), d = -0.301	-0.301	77	84	Between-subjects

Note. Positive effect sizes indicate that variety reduced perceived quantity.

Figure 10

Effect Sizes (Cohen's d) from the Original Studies, Replication Studies, and Average Effect Sizes with 95% Confidence Intervals



So, does variety reduce perceived quantity? We do not think so. We do not accuse Redden and Hoch of questionable research practices, but we doubt that the effect is robust. We hope that our research can guide future research by showing which unsuccessful replication studies have already been attempted regarding the effect of variety on perceived quantity.

CRediT Author Statement

- LR: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing
- GF: Conceptualization, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Writing – review & editing
- JA: Conceptualization, Formal analysis, Investigation, Methodology, Writing – review & editing
- AS: Supervision, Writing – review & editing

Author Contact

Correspondence concerning this article should be addressed to Lukas Röseler, lukas.roeseler@uni-muenster.de.

ORCID IDs: LR: 0000-0002-6446-1901, JA: 0000-0003-3846-2959, AS: 0000-0002-6358-167X

Conflict of Interest and Funding

The authors declare that they have no conflict of interest. This research was partly funded by a graduate scholarship granted by the state of Saxony-Anhalt, Germany, to Lukas Röseler and by the Harz University of

Applied Sciences. The funding sources were not involved in the study design, data collection, data analysis, interpretation of data, writing, or decision to submit.

Acknowledgments

We thank Theresa Bärthel, Anna Hofmann, Melina Meier, Sara Piske, Ina Siebert, Sabrina Vogt, Marcel Lüdtke, Ronja Pfeiffer, Miriam Dreher, Duy Anh Bui Thanh, Martina Sandforth, Melina Naumann, and Sarah Specht for their help in conducting the replication studies. We thank Joe Redden for ideas about why we could not replicate the effect and Jane Zagorski for language editing. We thank Tom Wallis, Matthew Inglis, and Aaron Charlton for their reviews and constructive criticism.

Open Science Practices

This article earned the Open Data, Open Materials, and Open Code badge for making the data, materials, and code openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement.

References

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). *How to perform a meta-analysis with R: A practical tutorial* [Computer software].

Burr, D. C., Anobile, G., & Arrighi, R. (2017). Psychophysical evidence for the number sense. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *373*(1740). https://doi.org/10.1098/rstb.2017.0045

Dakin, S. C., Tibber, M. S., Greenwood, J. A., Kingdom, F. A. A., & Morgan, M. J. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(49), 19552–19557. https://doi.org/10.1073/pnas.1113195108

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPower3-BRM-Paper.pdf

Frith, C. D., & Frith, U. (1972). The solitaire illusion: An illusion of numerosity. *Perception & Psychophysics*, *11*(6), 409–410. https://doi.org/10.3758/BF03206279

Ginsburg, N. (1978). Perceived numerosity, item arrangement, and expectancy. *The American Journal of Psychology*, *91*(2), 267. https://doi.org/10.2307/1421536

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). *dmetar: Companion R package for the guide 'doing meta-analysis in R'* (Version 0.0.9000) [Computer software]. http://dmetar.protectlab.org

Inquisit 3 [Computer software]. (n.d.). https://www.millisecond.com/download/archives.aspx

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kahn, B. E., & Wansink, B. (2004). The influence of assortment structure on perceived variety and consumption quantities. *Journal of Consumer Research*, *30*(4), 519–533. https://doi.org/10.1086/380286

Lumley, T., & Gordon, M. (2019). *forestplot: Advanced forest plot using 'grid' graphics* (Version 1.9) [Computer software]. https://CRAN.R-project.org/package=forestplot

Lutsch, C. (2001). *Equip Questionnaire Generator* [Computer software]. https://www.talvaro.com/

R Core Team. (2018). *R* [Computer software]. Vienna, Austria. https://www.R-project.org/

Redden, J. P., & Hoch, S. J. (2009). The presence of variety reduces perceived quantity. *Journal of Consumer Research*, *36*(3), 406–417. https://doi.org/10.1086/598971

Revelle, W. (2018). psych: Procedures for personality and psychological research [Computer software]. *Northwestern University, Evanston, Illinois, USA*, *1.8.3*. https://CRAN.R-project.org/package=psych

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows pre-

senting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution [Advance online publication]. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2160588

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science (New York, N.Y.)*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vos, P. G., van Oeffelen, M. P., Tibosch, H. J., & Allik, J. (1988). Interactions between area and numerosity. *Psychological Research*, *50*(3), 148–154. https://doi.org/10.1007/BF00310175

Wansink, B., & van Ittersum, K. (2003). Bottoms up! the influence of elongation on pouring and consumption volume. *Journal of Consumer Research*, *30*(3), 455–463. https://doi.org/10.1086/378621

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H., & Bryan, J. (2018). *readxl: Read excel files* (Version 1.1.0) [Computer software]. https://CRAN.R-project.org/package=readxl

Wolf, D. (2021). *Files* [Computer software]. https://osf.io/zau4r

Yeager, D. S., Bryan, C., & O'Brien, J. (2019). Replicator degrees of freedom allow publication of misleading 'failures to replicate' [Advance online publication]. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3408200

Zhao, J., & Yu, R. Q. (2016). Statistical regularities reduce perceived numerosity. *Cognition*, *146*, 217–222. https://doi.org/10.1016/j.cognition.2015.09.018