Do Open Science Badges Increase Trust in Scientists among Undergraduates, Scientists, and the Public?

Running head

Do Open Science Badges Increase Trust in Scientists?

Authors

Jürgen Schneider[a]	juergen.schneider@uni-tuebingen.de	0000-0002-3772-4198
Tom Rosman[b]	tr@leibniz-psychology.org	0000-0002-5386-0499
Augustin Kelava[c]	augustin.kelava@uni-tuebingen.de	0000-0001-6053-0415
Samuel Merk[d]	samuel.merk@ph-karlsruhe.de	0000-0003-2594-5337

[a] University of Tübingen, School of Education, Tübingen, Germany

[b] Leibniz Institute for Psychology Information, Department of Research Literacy and User Friendly Research Support, Trier, Germany

[c] University of Tübingen, Methods Center, Tübingen, Germany

[d] University of Education Karlsruhe, Institute for School and Instructional Development in Primary and Secondary Education, Karlsruhe, Germany

Corresponding Author

Jürgen Schneider

juergen.schneider@uni-tuebingen.de

+49 7071 29 73617

Keplerstr. 17, 72074 Tübingen, Germany

Do Open Science Badges Increase Trust in Scientists among Undergraduates, Scientists, and the Public?

Abstract

Open science badges are a promising method to signal a study's adherence to open science practices (OSP). In three experimental studies, we investigated whether badges affect trust in scientists by undergraduates ($N = 270$), scientists ($N = 250$), or the public ($N = 257$). Furthermore, we analyzed the moderating role of epistemic beliefs in this regard. Participants were randomly assigned to two of three conditions: Badges awarded (visible compliance to OSP), badges not awarded (visible noncompliance to OSP), and no badges (control). In all samples, our Bayesian analyses indicated that badges influence trust as expected with one exception in the public sample: an additional positive effect of awarded badges compared to no badges was not supported here. Further, we found evidence for the absence of a moderation by epistemic beliefs. Our results demonstrate that badges are an effective means to foster trust in scientists among target audiences of scientific papers.

Keywords

badges, trust, epistemic beliefs, open science, open data, open materials, preregistered

Statement of Relevance

Open science practices (such as open data, open materials, or open code) are increasingly being called for, not only in psychological science but in all disciplines involving empirical methods. Several journals currently use badges to signal the compliance of specific articles to open science practices and foster incentive structures for transparency. To our knowledge, our study is the first to investigate how these badges affect individual factors such as trust and epistemic beliefs. We find that these badges increase trust in scientists and reduce multiplistic epistemic beliefs of undergraduates and scientists. Our research thus contributes to the evidence that badges "work," which will likely further incentivize researchers' commitment to open science practices. Furthermore, our results on epistemic beliefs indicate that badges may help to promote an idea of science that is not just an "opinion."

Do Open Science Badges Increase Trust in Scientists among Undergraduates, Scientists, and the Public?

In recent times, struggles in replicating empirical findings has been acknowledged by several scientific disciplines (Camerer et al., 2018; Open Science Collaboration, 2015). Recent studies support the assumption of a detrimental effect of this so-called replication crisis on perceived trustworthiness (Anvari & Lakens, 2018; Wingen et al., 2020). A primary reaction to this was the call for scientists to increase the transparency and reproducibility of the entire research process (Lindsay, 2015; Vazire, 2018). To signal adherence to open science practices (OSP), a number of academic journals have adopted open science badges, which allow quickly determining whether a study has implemented OSP—an important indicator for gauging its transparency and trustworthiness. However, besides some first indications of their effectiveness to foster the implementation of OSP (Kidwell et al., 2016), not much is known on the effects of badges at an individual level. Therefore, we investigated in three studies how trustworthy scientists are perceived by undergraduates, scientists, or the public, depending on the inclusion of badges in their articles. Furthermore, considering the crucial role of beliefs about science in information processing, we explore the potential role of epistemic beliefs in moderating the effectiveness of badges and indirectly predicting trust itself.

Epistemic trust

In our closely connected world, which is characterized by the division of cognitive labor, we are dependent on other people's knowledge (Bromme, Kienhues & Prosch, 2010). However, we cannot evaluate the truthfulness of all information from sources we interact with, particularly when lacking resources for judgment such as knowledge, time and financial capital (Stadtler & Bromme, 2014; Zimmermann & Jucks, 2018). Recipients of scientific claims usually have limited access to first-hand information (e.g., the concrete research process), since they are not involved in the research process itself (Bromme & Goldman, 2014; Hendriks & Kienhues, 2019). Journal articles mostly summarize the underlying research process, and press releases or translational abstracts ("plain language summaries") often only provide overviews. Consequently, readers of scientific claims cannot evaluate the truthfulness of a scientific claim by themselves but have to rely (to various degrees) on so-called second-hand evaluations (i.e., evaluations on the trustworthiness of an information source instead of

the information itself; Bromme et al., 2010). Therefore, when acquiring and evaluating information, trust plays a pivotal role, as shown in studies on decision making (Isen, 2008; Liu, Vanderbilt, & Heyman, 2013) and learning (Landrum, Eaves, & Shafto, 2015). This is equally true for different population groups, each interacting with scientific claims from their specific perspectives: Scientists in their daily work, undergraduate students in their professional development (e.g., student teachers, Munthe & Rogne, 2015), and the public through science communication (e.g., on public health recommendations during a pandemic, Andrews Fearon, Götz, & Good, 2020).

On a conceptual level, we define *trust* as beliefs about the trustee's characteristics that make him or her favorable toward the trustor and consequently vulnerable to actions of the trustee (McCraw, 2015). Research syntheses on the topic of trust (e.g., Mayer, Davis, & Schoorman 1995) particularly highlight benevolence, integrity, and expertise as dimensions of trust (or, closely related, competence and warmth; Fiske, Cuddy, & Glick, 2007). More specifically, *epistemic* trust addresses the development and justification of knowledge (Origgi, 2014), as is the case with research reports on evidence generated by scientists.

Open science practices and epistemic trust

Alongside goals of research quality and development (Fecher & Friesike, 2014), researchers exposing themselves to scrutiny by disclosing their scientific practices may help to rebuild trust in scientists (Grand, Wilkinson, Bultitude & Winfield, 2012) as it signals integrity on the part of the (trusted) researcher (Lyon, 2016). In line with these assumptions, a recent U.S. survey reveals that adults would trust scientific research findings more if the corresponding data were openly available (Pew Research Center, 2019). Recently, these findings were corroborated for the [removed for blind peer-review] context by Author et al. (in preparation). Furthermore, Soderberg, Errington and Nosek (2020) report similar results on credibility judgments by scientists about preprints: Participants indicated the availability of research materials, data, and data analysis scripts as the most relevant factor for their judgements.

These findings raise the question of how to signal adherence to OSP effectively. Recent research suggests that, at least for the general public, a straightforward communication strategy is not enough to rebuild trust in past research (Wingen et al., 2020) and it may even further decrease trust in future

research (Anvari & Lakens, 2018). The interventions used in these studies implemented rather decontextualized descriptions of OSP on a discipline-specific level. Therefore, participants not familiar with the scientific process (i.e., nonscientists with low scientific literacy) might not fully comprehend how these rather abstract "reforms" shape research practice (Laugksch, 2000) and why they consequently might help to improve replicability. Furthermore, the reforms were communicated as set goals or statements of intent (e.g., "Based on this problematic result, psychological researchers now aim to make their research more transparent," Wingen et al., 2020, supplemental material). These may appear less convincing to participants than actual implementations of the reforms that have been certified by third parties (Chang, Cheung & Tang, 2013).

In our view, badges are a more tangible and contextualized way to signal the attested adherence to or violation of standards concerning certain aspects of OSP (Bauer, 2020). Academic journals have increasingly adopted the practice of awarding OSP badges in recent years (for a listing of practicing journals, see https://www.cos.io/our-services/badges). Initial investigations indicate that badges are related to a higher frequency of OSP and a better adherence to OSP standards, particularly concerning data sharing (Kidwell et al., 2016). We, therefore, argue that the badges displayed on scientists' publications or translational abstracts influence the perceived trustworthiness of the authors, with colored badges, signaling the adherence to qualitative standards, increasing trust and grayed out badges signaling the violation of qualitative standards, decreasing trust compared to no badges.

Hypothesis 1 (H1): Visible compliance to OSP (colored badges) lead to higher perceived trustworthiness of scientists compared to no information about OSP (control condition) or visible noncompliance to OSP (grayed out badges), with visible noncompliance to OSP receiving the lowest ratings of trustworthiness.

Epistemic beliefs and epistemic trust

Epistemic beliefs—individual beliefs about the nature of knowledge and knowing (Hofer & Pintrich, 1997) —are known to influence information processing when dealing with textual information (Bråten, Britt, Strømsø & Rouet, 2011; Franco et al., 2012). Developmental conceptualizations of epistemic beliefs distinguish between the consecutive stages of absolutism (knowledge as dualistic, "right-or-wrong"), multiplism (knowledge as subjective opinions), and evaluativism (knowledge as

weighed evidence). Because of their focus on personal opinions over facts and evidence, multiplistic beliefs (Kuhn & Weinstock, 2002) seem to impair information processing in particular – as evidenced by their negative effects on learning (Author et al., 2018a) and negative relationships with judgments of text trustworthiness (Strømsø et al., 2011).

Hypothesis 2 (H2): The higher the multiplistic beliefs, the lower the perceived trustworthiness of scientists.

Furthermore, multiplistic beliefs depict the source of knowledge as something that lies within a knowing subject in the form of individual opinions. Individuals with high levels of multiplistic beliefs thus perceive external sources of knowledge (e.g., researchers) and knowledge evaluation (e.g., through badges) as irrelevant because they consider all knowledge claims to be equally true (Kuhn & Weinstock, 2002). For these individuals, the question of how knowledge from external sources is created or displayed may therefore be unrelated to their perceptions of trustworthiness. Consequently, we assume that for individuals with high levels of multiplistic beliefs, badges will not play a role regarding their epistemic trust. Since no corresponding empirical evidence exists to date, we, however, label this hypothesis as exploratory.

Hypothesis 3 (H3): Multiplistic epistemic beliefs moderate the effect of badges on perceived trustworthiness.

Moreover, badges might indicate that science is not just "opinion" because they make the underlying empirical and fact-based approach more tangible, thus reducing multiplistic beliefs. We, however, concede that this interpretation is somewhat speculative, which is why we, again, label the corresponding hypothesis as exploratory.

Hypothesis 4 (H4): Visible compliance to OSP (colored badges) lead to lower multiplistic epistemic beliefs compared to no information about OSP (control condition) or visible noncompliance to OSP (grayed out badges).

Study 1: Undergraduate Students

In the first study, we investigated our research questions in a sample of student teachers. Participants from this population regularly access scientific papers in the course of their professional development,

as evidence-based practice plays a central role in [removed for blind peer-review] teacher education curricula (Cochran-Smith, 2009).

Method

Design

Hypotheses were tested in an experiment with three conditions: Students were presented two title pages of fictitious empirical journal articles (topics: dual channel theory, learning by means of worked out examples), whereby these title pages contained either (a) three colored badges with legends (condition "colored badges", CB), or (b) three grayed out badges with legends (condition "grayed out badges", GB), or (c) no badges ("control condition", CC), but also legends which explained other terms on the title page (see Figure 1). The three colored badges indicated that the authors implemented the open science practices "open data," "open materials," and "open code," and the grayed out badges signaled nonadherence with these practices "data not available," "materials not available," and "code not available." As we expected participants to not be familiar with badges, we included explanations of the badges in gray text boxes (see Figure 1). These were explicitly labeled as additional information that was not part of the journal article itself. In the condition without badges, participants did not receive information about the implementation of OSP. To prevent experimental leakage, but at the same time increase test power, we used a planned missing design (Graham, Cumsille & Elek-Fisk, 2003; Silvia, Kwapil, Walsh & Myin-Germeys, 2014): Each participant completed two of the three conditions. A balanced experimental plan was used to randomize the assignment and sequence of conditions, as well as the topics and sequence of topics.

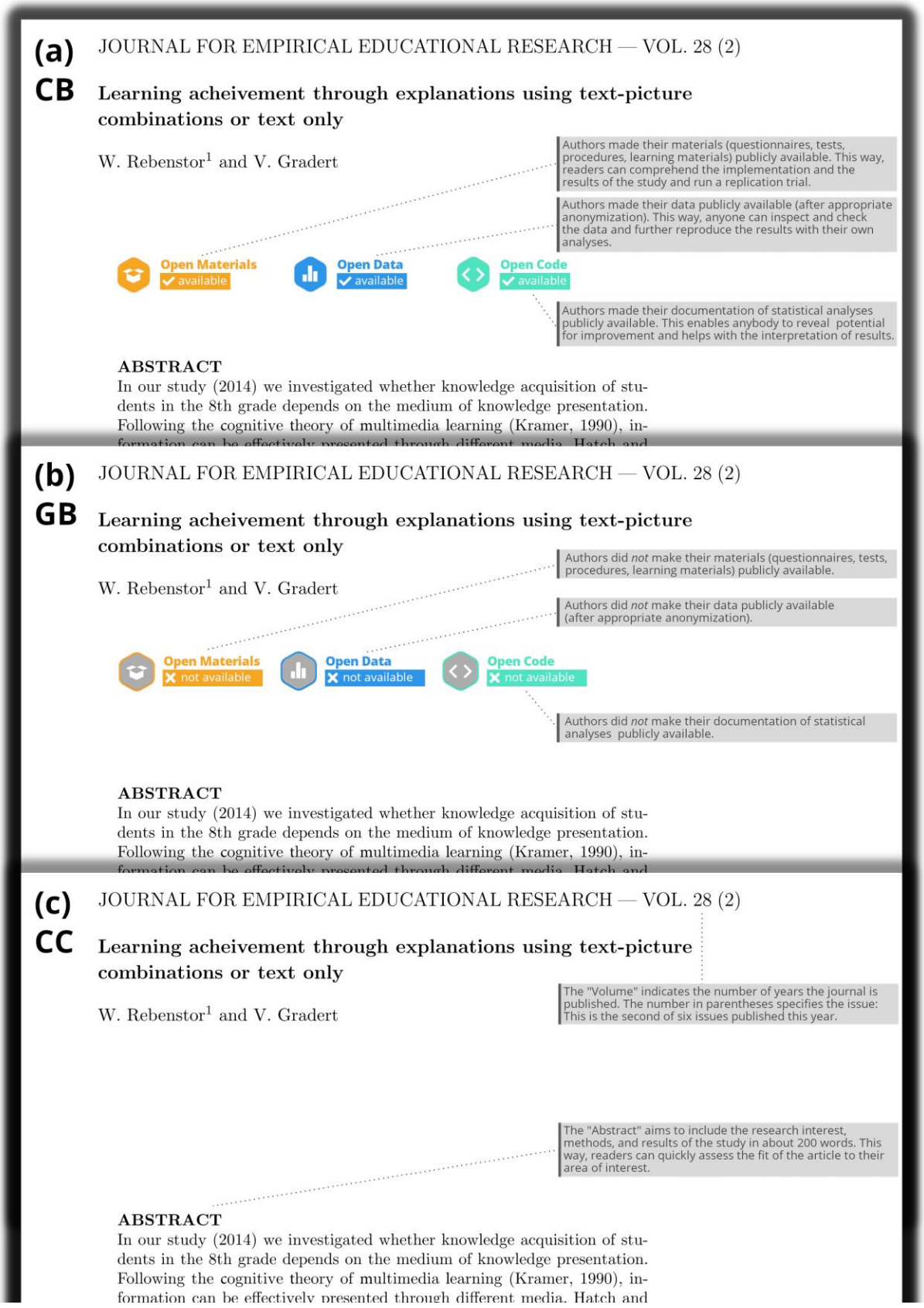

Figure 1

Illustrations of the three experimental conditions (upper part of the title pages). (a) CB: Colored badges, (b) GB: grayed out badges, (c) CC: control condition

Procedure

After participants gave their informed consent, they were introduced to the survey procedure and informed about its structure. They were told that they would be given the title page of a regular journal article with explanations annotated in gray text boxes. Participants were asked to read the title page thoroughly and then answer the questions below the text. On the next survey page, participants read the title page of the first journal article and were prompted to respond to a topic-specific multiplism scale (see below; Author et al., 2018b). Subsequently, they completed the Muenster Epistemic Trustworthiness Inventory (METI, Hendriks, Bromme & Wicherts, 2015), and the treatment check was conducted. This sequence of events was repeated for the second title page. Finally, at the end of the questionnaire, participants responded to several demographic questions. The survey took approximately 15 minutes to complete (for a demo version of the survey with all three conditions visit https://undergrad-demo.formr.org).

Statistical Analyses

For data analyses we used the (approximate adjusted fractional) Bayes factors (Gu, Mulder & Hoijtink, 2018; Hoijtink, Mulder, van Lissa & Gu, 2019) for informative hypotheses, as they are especially suitable to test hypotheses with order restrictions (Hoijtink, 2012), like ours. To ensure a strictly confirmatory approach (Wagenmakers, Wetzels, Borsboom, van der Maas & Kievit, 2012), we preregistered our hypotheses [removed for blind peer-review, reviewers: see file attached]. Within this preregistration, we specified a data analysis plan which, in turn, served as a basis for our simulation-based sample size determination (Bayes factor design analysis, see Schönbrodt & Wagenmakers, 2018). This data analysis strategy and the results of the sample size determination are described in the following.

Bayes factors, in general, provide relative evidence as they quantify the increased likelihood that the current data are observed under a specific hypothesis in contrast to a different hypothesis. Therefore, a central challenge is choosing which hypotheses to compare to each other in order to gain the most

compelling evidence. Our first hypothesis, H1, stated that student teachers would ascribe, on average, less integrity (int) to the authors of studies if these title pages contained grayed out badges (GB) compared to title pages with no information about the use of OSP (CC), which, in turn, would be ascribed less integrity than authors of title pages containing colored badges (CB). In our preregistration we specify comparing this hypothesis $H1_1$: $\mu(int)_{GB} < \mu(int)_{CC} < \mu(int)_{CB}$ with the corresponding point null-hypothesis $H1_0$: $\mu(int)_{GB} = \mu(int)_{CC} = \mu(int)_{CB}$ and a hypothesis that assumes that only the visible adherence to OSP has an effect on integrity $H1_2$: $\mu(int)_{GB} = \mu(int)_{CC} < \mu(int)_{CB}$. Furthermore, in our preregistration, we specify that if the data provide evidence for one of these hypotheses against the other two (Bayes Factor: BF > 3 respective < 1/3) and the corresponding hypothesis without constraints $H1_u$: $\mu(int)_{GB}$; $\mu(int)_{CC}$; $\mu(int)_{CB}$, we would compare this hypothesis to its complement $\overline{H1_i}$ (which contains all mean configurations that do not satisfy the restrictions of $H1_i$). Only when all these comparisons also result in Bayes factors outside the interval [1/3; 3], do we consider our results as evidence for $H1_i$ and otherwise as inconclusive.

We computed these Bayes factors using the routines implemented in the R package bain (Gu, Hoijtink, Mulder & Rosseel, 2019). This statistical package uses an adjusted and approximated version of the fractional Bayes factor, which, in turn uses a fraction of the information in the data to specify the implicit prior (for details, see Gu et al., 2018). This framework is especially useful for our analyses, as it provides a routine for computing Bayes factors using multiple imputation data (Hoijtink, Gu, Mulder & Rosseel, 2019). Correspondingly, we imputed our (planned as well as unplanned) missing data using chained equations (Azur, Stuart, Frangakis & Leaf, 2011; van Buuren, 2012). Next, parameters of a repeated measurement ANOVA were estimated on each of the resulting (1,000) complete data sets and combined using the rules derived by Hoijtink, Gu, et al. (2019).

To determine our preregistered sample size, we ran simulation studies that used the decision procedure described above and assumed Cohen's $d = .3$ if $\mu(int)_x \neq \mu(int)_y$. The simulations suggested that a sample size of 250 would be sufficient, as, in the worst case (true hypothesis is $H1_2$), our decision procedure would result in evidence for an incorrect hypothesis in only 2% of the simulated cases and would remain inconclusive in 28% of the simulated cases (see preregistration for details).

Sample

According to the preregistration, we started recruiting the sample by advertising in social media groups and newsletters for student teachers from [removed for blind peer-review] universities. According to our stopping rule, we stopped data collection at $N = 270$. We exceeded the stopping rule by $n = 20$ participants, as the survey had to be deactivated manually after periodic sample size checks. Thirteen participants skipped the repeated measurement, and four did not complete the demographic questions at the end of the questionnaire. On average, participants were 22.89 years ($SD = 2.95$) and in their sixth semester ($M = 5.86$, $SD = 3.68$). Of all the participants, 176 indicated female gender.

Instruments

All studies were conducted using the web-based survey tool *formr* (Arslan, Walther & Tata, 2020).

Integrity

The METI (Hendricks et al., 2015) was used to assess the degree of integrity participants ascribed to the authors of the respective title page. This instrument contains 14 antonym pairs that are rated on a 7-point scale and are mapped to three subscales (expertise: *well educated–poorly educated*; integrity: *honest–dishonest*; benevolence: *considerate–inconsiderate*). Even though we were only interested in one dimension of the inventory (see preregistration), participants completed all three dimensions because we wanted to gain some additional insights on the instrument's construct validity and to use the additional information as covariates to impute the planned missing data. Therefore, we first performed a confirmatory factor analysis (CFA) with τ-congeneric measurement models for each measurement, which resulted in good fit indices (see Table 1) after freeing two residual covariances. In a next step, we further investigated the factorial structure using a two-level confirmatory factor analysis (CFA), and its good model fit corroborated the assumption of three dimensions at the within-person as well as at the between-person levels (see Table 1 and the reproducible documentation of the analysis [RDA] for details). Furthermore, all three-dimensional models significantly outperformed corresponding one-dimensional models (p-values of χ^2 difference test are all smaller than .0001). As we specified τ-congeneric measurement models, McDonald's ω was used to assess internal

consistency (Dunn, Baguley & Brunsden, 2014), and this yielded good results, with a minimum score of ω = .83 (integrity in the first measurement).

Topic specific multiplism

To assess topic specific multiplism, we used a 4-point Likert-type scale by Author et al. (2018b, sample item: "*The insights from the text are arbitrary*"). Consecutive as well as two-level CFAs provided evidence for the assumption of one-dimensionality (see Table 1), and the scale's internal consistency was acceptable considering its length (four items, ω = .65 and ω = .53 for the topics respectively).

Treatment check

To investigate the effectiveness of our treatment, we examined whether participants recognized and understood the presented badges. To do so, we, directly and indirectly, asked them about their perceptions of the researchers' OSP (five 4-point Likert-type items with a "don't know" option, e.g.: *Materials used in the study and the data collected are openly accessible.* 1 = *I do not agree at all*, 4 = *fully agree*). A corresponding CFA yielded excellent results (see Table 1), and the internal consistency of the treatment check was also very good (ω = .95 and ω = .90).

Table 1

Results of the CFAs with Fit Indices (Study 1)

	1d CFA METI 1	1d CFA METI 2	3d CFA METI 1	3d CFA METI 2	1d MCFA METI	3d MCFA METI	1d CFA TSM 1	1d CFA TSM 2	1d MCFA TSM	1d CFA TCH 1	1d CFA TCH 2
χ^2	588.932	936.555	194.174	212.262	764.777	277.759	5.302	6.072	4.422	2.412	6.538
df	77.000	77.000	73.000	72.000	154.000	145.000	4.000	4.000	4.000	3.000	3.000
CFI	.811	.759	.955	.961	.894	.977	.991	.990	.999	1.000	0.997
TLI	.776	.715	.944	.950	.875	.971	.987	.985	.996	1.003	0.991
RMSEA	.157	.208	.078	.087	.087	.042	.035	.045	.014	0.000	0.083
SRMR	.084	.099	.049	.040	.271	.172	.048	.055	.109	0.006	0.005
SRMR between	-	-	-	-	.146	.091	-	-	.090	NA	NA
SRMR within	-	-	-	-	.125	.081	-	-	.019	NA	NA
BIC	10225.876	9336.494	9853.512	8639.946	18914.759	18484.147	2791.127	2653.377	5445.587	1547.731	1750.473
AIC	10125.120	9237.119	9738.362	8522.826	18616.055	18147.038	2769.537	2632.082	5360.243	1512.868	1712.703

Note: 1d: one-dimensional, 3d: three-dimensional, METI: Muenster Epistemic Trustworthiness Inventory, TSM: topic specific multiplism, TCH: treatment check

Results

Treatment check

Figure 2 depicts a fluctuation diagram (also known as "product plot," Wickham & Hofmann, 2011) of the results of the treatment check. We consider these results as evidence for strong compliance with our treatment, as, for example, comparing the conditions GB and CB resulted in large effect sizes for ordinal measures (e.g., Varha & Delaney's A = .84 for Item 1). In the CC condition, a high proportion

of participants reported not knowing about the researchers' OSP or their judgments showed high variation.

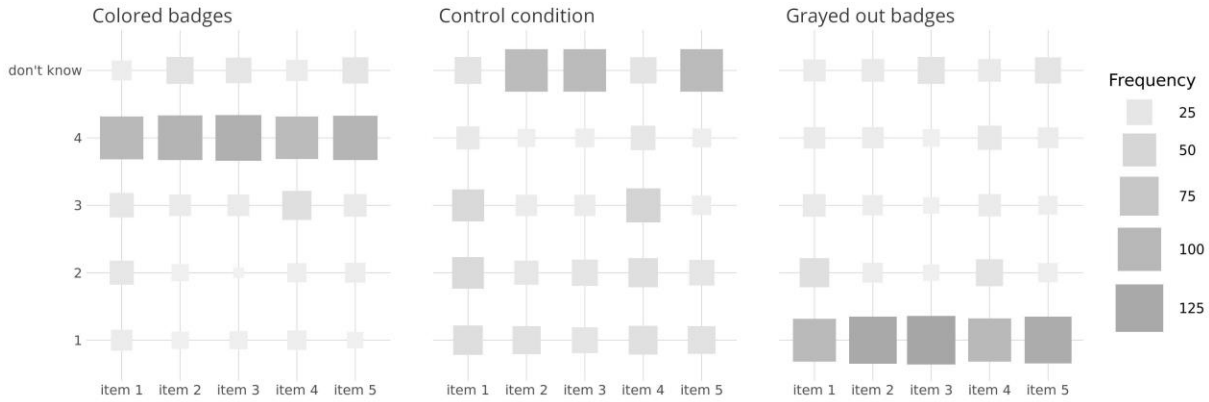

Figure 2

Fluctuation diagram of the results from the treatment check in Study 1. Frequency per item and experimental condition.

Hypothesis 1

H1 states that the CB condition induces higher perceived integrity of the authors than the CC, which, in turn, induces higher perceived integrity than the GB condition. To test H1, we applied the preregistered equation to compute the approximate adjusted fractional Bayes factors for the corresponding Hypothesis $H1_1$: $\mu(int)_{GB} < \mu(int)_{CC} < \mu(int)_{CB}$, the point null-hypothesis $H1_0$: $\mu(int)_{GB} = \mu(int)_{CC} = \mu(int)_{CB}$, and a hypothesis that postulates only an effect of the visible utilization on integrity $H1_2$: $\mu(int)_{GB} = \mu(int)_{CC} < \mu(int)_{CB}$, whereby $\mu(int)_X$ describes the mean of integrity in the group X (see statistical analysis section). As the underlying ANOVA model for such hypotheses assumes normality of the dependent variable, we first checked if the data satisfied this assumption regarding skewness, kurtosis, and outliers. As the data showed no strong violations of these criteria, we continued by (multiply) imputing the planned and unplanned missing data using the procedures implemented in the mice package for R (van Buuren & Groothuis-Oudshoorn, 2011). Using this data, we followed the preregistered decision procedures previously described in the statistical analyses section. This resulted in substantial relative evidence for $H1_1$ (BF against $H1_0 = 3.5 \cdot 10^7$, BF against

$H1_2$ = 4.5·10[1], BF against $\overline{H1_1}$ = 4.8·10[3], BF against $H1_u$ = 5.5). Furthermore, comparing the means of integrity between the three experimental groups resulted in moderate to large effect sizes, $d_{GB/CC}$ = .32, $d_{CC/CB}$ = .29, and $d_{GB/CB}$ = -.57 (see Figure 3).

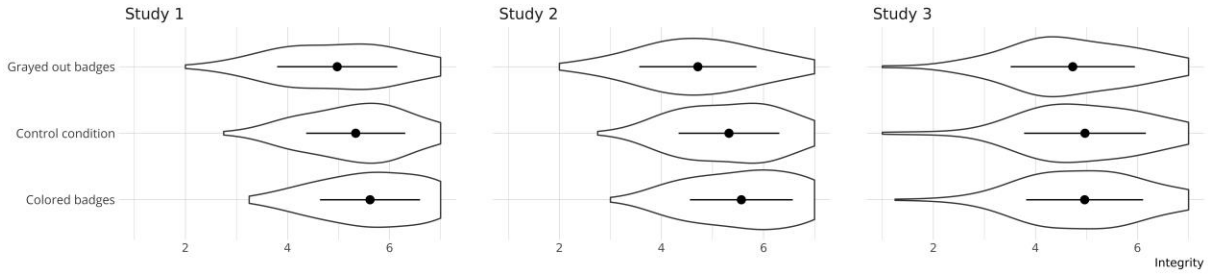

Figure 3

Violin plots and means ± 1SD of integrity by experimental condition for all three studies.

Hypothesis 2

H2 predicted a negative association between topic specific multiplism and integrity. To test this hypothesis, we specified a path model with three regression paths—one for each condition of topic specific multiplism on integrity (see Figure 4). Subsequently, we tested the hypothesis $H2_1$: $b_1^{CB} > 0$ & $b_1^{CC} > 0$ & $b_1^{GB} > 0$ against $H2_0$: $b_1^{CB} = 0$ & $b_1^{CC} = 0$ & $b_1^{GB} = 0$, again using the approximate adjusted fractional Bayes factor, which resulted in strong evidence for $H2_1$ (BF against $H2_0$ = 6.0·10[21], BF against $\overline{H2_1}$ = 2.4·10[7], BF against $H2_u$ = 6.3). Figure 4 depicts the pooled standardized regression coefficients as a measure of effect size.

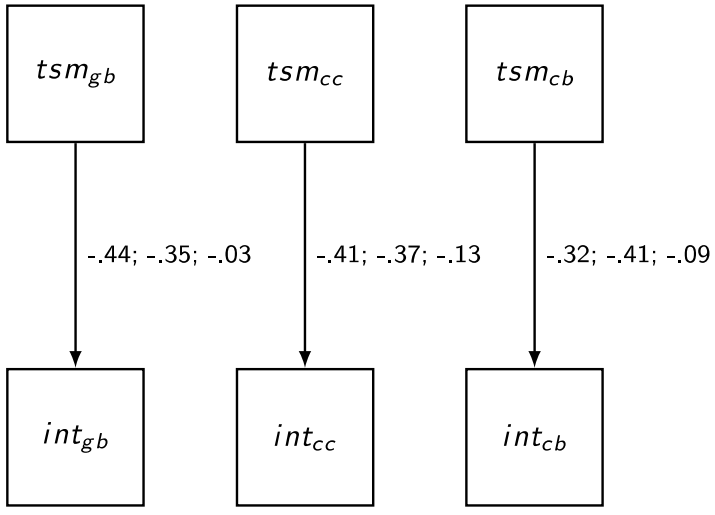

Figure 4

Path model for H3 and H4 with pooled estimates of all three studies (samples of undergraduates, social scientists, and the public). For visual clarity, we did not depict variances and covariances. tsm = topic specific multiplism, int = integrity, gb = grayed out badges, cc = control condition, cb = colored badges.

(Exploratory) Hypothesis 3

Figure 4 also shows the results obtained for H3, which states that the association between topic specific multiplism and integrity may be moderated by the topic, resulting in the following order of $H3_1$: $b_1^{GB} > b_1^{CC} > b_1^{CB}$. We tested this hypothesis against the corresponding null hypothesis $H3_0$: $b_1^{GB} = b_1^{CC} = b_1^{CB} = 0$ and a hypothesis which states $H3_2$: $(b_1^{GB}, b_1^{CC}) > b_1^{CB}$, meaning that the association is smaller when participants were informed about the use of open science practices, but every configuration between the other coefficients is allowed. The Bayes factors clearly provided relative evidence for the null hypothesis (BF against $H3_1$ = 6.0, BF against $H3_2$ = 7.4, BF against $\overline{H3}_0$ = 18.5, BF against $H3_u$ = 18.5).

(Exploratory) Hypothesis 4

Finally, we tested if the condition also had an effect on topic specific multiplism. The violin plots depicted in Figure 5 indicate that there might be small to moderate effects. This is underpinned by the effect size estimates ($d_{GB/CC}$ = -.26, $d_{CC/CB}$ = .01, $d_{GB/CB}$ = -.25) and the Bayes factors which favor $H4_1$: $\mu(tsm)_{GB} > \mu(tsm)_{CC} > \mu(tsm)_{CB}$ against a corresponding null hypothesis $H4_0$: $\mu(tsm)_{GB} = \mu(tsm)_{CC} = \mu(tsm)_{CB}$ and a less specific hypothesis $H4_2$: $(\mu(tsm)_{GB}, \mu(tsm)_{CC}) > \mu(tsm)_{CB}$, which only states that topic specific multiplism is smaller when participants are confronted with OSP badges (BF against $H4_0$ = 6.2, BF against $H4_2$ = 1.9, BF against $\overline{H4}_1$ = 8.4, BF against $H4_u$ = 3.6).

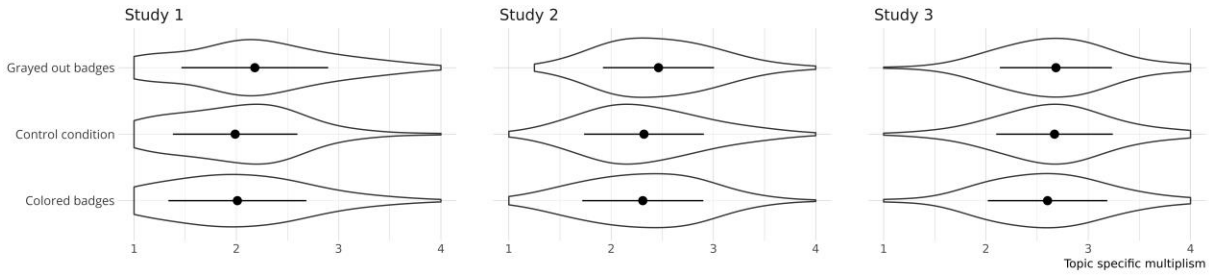

Figure 5

Violin plots and means ± 1SD of topic specific multiplism by experimental condition for all three studies.

Study 2: Social Scientists

In a second study, we aimed to replicate the findings from the first study in a sample of social scientists. This sample is expected to be more practiced in working with publications and possibly have more knowledge of OSP badges.

Method

Design

The design of the conditions was the same as in Study 1. To avoid potential bias in the participants' judgments due to topic familiarity (Tversky & Kahneman, 1973), abstracts of fictional studies were used (see supplemental material). In a small-scale pilot study ($N = 39$), we tested and confirmed the authenticity of these abstracts. We implemented the abstracts in the design of the title pages from Study 1. Again, the same experimental conditions (CB, CC, GB) were realized. We also used the same planned missing design and assigned participants randomly to the different conditions using a balanced experimental plan.

Procedure and Statistical Analyses

All procedures and statistical analyses were the same as in Study 1. For a demo version of the survey with all three conditions, visit https://sci-demo.formr.org.

Instruments

Participants completed the same instruments as in Study 1. Internal consistency was very good for *integrity* (ω = .91 and ω = .92), acceptable for *topic specific multiplism* (ω = .69 and .64), and very good for the *treatment check* (ω = .87 and .91).

Table 2

Results of the CFAs with fit indices (Study 2)

	1d CFA METI 1	1d CFA METI 2	3d CFA METI 1	3d CFA METI 2	1d MCFA METI	3d MCFA METI	1d CFA TSM 1	1d CFA TSM 2	1d MCFA TSM	1d CFA TCH 1	1d CFA TCH 2
χ^2	463.608	479.509	171.668	166.600	473.759	238.803	5.245	7.677	1.023	9.287	0.269
df	77.000	77.000	74.000	74.000	154.000	148.000	4.000	4.000	2.000	4.000	2.000
CFI	0.881	0.896	0.970	0.976	0.954	0.987	0.993	0.974	1.000	0.996	1.000
TLI	0.860	0.878	0.963	0.971	0.945	0.984	0.990	0.962	1.020	0.989	1.006
RMSEA	0.142	0.145	0.073	0.071	0.064	0.035	0.035	0.061	0.000	0.089	0.000
SRMR	0.057	0.053	0.029	0.029	0.304	0.125	0.034	0.058	0.027	0.008	0.001
SRMR between	-	-	-	-	0.232	0.085	-	-	0.019	-	-
SRMR within	-	-	-	-	0.072	0.040	-	-	0.008	-	-
BIC	8854.815	8235.105	8579.439	7938.761	16735.574	16537.906	2369.599	2237.418	4559.431	1667.935	1685.730
AIC	8756.214	8136.504	8470.274	7829.596	16440.552	16217.596	2334.384	2216.290	4466.709	1633.572	1644.514

Note: 1d: one-dimensional, 3d: three-dimensional, METI: Muenster Epistemic Trustworthiness Inventory, TSM: topic specific multiplism, TCH: treatment check

Sample

As the social sciences predominantly utilize empirical methods in research, we opted for a social scientist sample. Participants were recruited via the online access panel provider prolific.co, filtering for social scientists. Following our stopping rule, we terminated data collection after $N = 250$

participants had passed the implemented quality check. No participant skipped the repeated measurement or the demographic questions at the end of the questionnaire. Ninety-one participants were younger than 35 years, 37 participants were between the ages of 35 and 49 years, and 20 were older than 50 years. Most participants described their current position as a graduate research assistant or postgraduate researcher (91). Female gender was indicated by 170 participants.

Results

Treatment check

As shown in Figure 6, Study 2 participants also complied very well with the treatment. The effect size for the first item comparing the CB and GB conditions was even larger than in Study 1 (Varga & Delaney's A = .94).

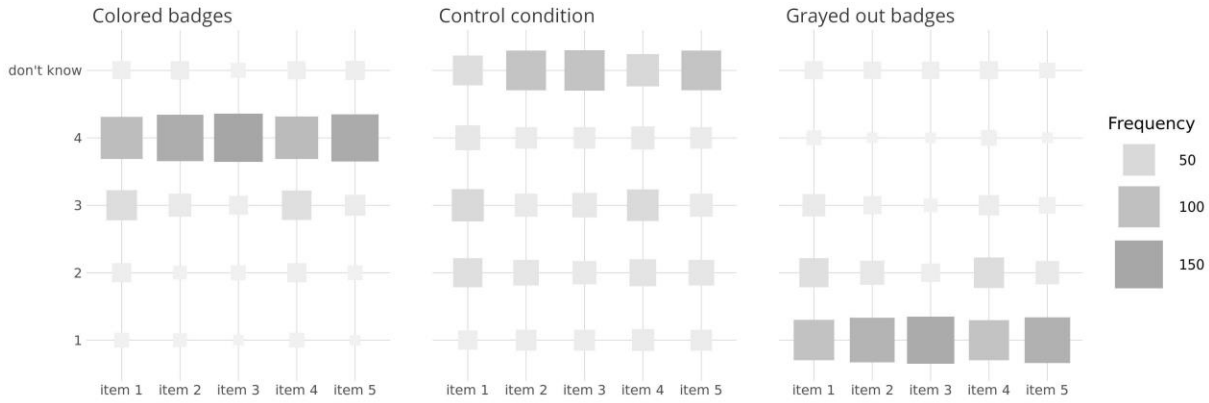

Figure 6

Fluctuation diagram of the results from the treatment check in Study 2. Frequency per item and experimental condition.

Hypothesis 1

Figure 3 already provides some insights with regard to $H1_1$: $\mu(int)_{GB} < \mu(int)_{CC} < \mu(int)_{CB}$. Following the same (preregistered) procedure as in Study 1, we again obtained substantial relative evidence for $H1_1$ (BF against $H1_0 = 1.6 \cdot 10^{11}$, BF against $H1_2 = 7.5$, BF against $\overline{H1}_1 = 4.8 \cdot 10^3$, BF against $H1_u = 5.4$) with moderate to large effect sizes, $d_{GB/CC} = .55$, $d_{CC/CB} = .25$, and $d_{GB/CB} = .77$.

Hypothesis 2

In Study 2, the results regarding H2 were also replicated: Testing the hypothesis $H2_1$: $b_1^{CB} > 0$ & $b_1^{CC} > 0$ & $b_1^{GB} > 0$ against $H2_0$: $b_1^{CB} = 0$ & $b_1^{CC} = 0$ & $b_1^{GB} = 0$ revealed strong evidence for $H2_1$ (BF against $H2_0 = 2.6 \cdot 10^{16}$, BF against $\overline{H2}_1 = 2.6 \cdot 10^{7}$, BF against $H2_u = 6.9$) with similar (moderate) effect sizes as in Study 1 (see Figure 4).

(Exploratory) Hypothesis 3

As in Study 1, the Bayes factors found for the exploratory H3 provided strong relative evidence for the null hypothesis (BF against $H3_1 = 1.3 \cdot 10^{2}$, BF against $H3_2 = 1.2 \cdot 10^{2}$, BF against $\overline{H3}_0 = 53.7$, BF against $H3_u = 53.7$).

(Exploratory) Hypothesis 4

Finally, Study 2 revealed very similar results to Study 1 with regard to H4, showing moderately higher means in topic specific multiplism for the condition with grayed out badges ($d_{GB/CC}$ = -.27, $d_{CC/CB}$ = .02, $d_{GB/CB}$ = -.24), which is reflected by Bayes factors clearly favoring $H4_1$: $\mu(tsm)_{GB} > \mu(tsm)_{CC} > \mu(tsm)_{CB}$ against a corresponding null hypothesis $H4_0$: $\mu(tsm)_{GB} = \mu(tsm)_{CC} = \mu(tsm)_{CB}$ (BF = 22.4), but not conclusively against the less specific alternative hypothesis $H4_2$: $(\mu(tsm)_{GB}, \mu(tsm)_{CC}) > \mu(tsm)_{CB}$ (BF = 2.0).

Study 3: General Public

Scientific findings also reach larger target groups, such as the general public, through science communication and science journalism. In the third study, we, therefore, aimed to replicate the findings from the two preceding studies in a sample of the general public.

Method

Design

Experimental conditions were identical to Studies 1 and 2. Additionally, the abstracts implemented on the title pages were adapted to the public's needs and levels of expertise. In the context of science

communication, authors are increasingly being asked to meet these needs and to promote the comprehension of research findings by laypeople (Author et al., 2021; Stricker, Chasiotis, Kerwer & Günther, 2020). Preparing translational abstracts is one approach endorsed by the American Psychological Association (APA; Kaslow, 2015). In addition to the scientific abstract accompanying scientific papers, the authors also prepare a translational abstract that is directed toward a public audience and free of technical language and scientific jargon. To illustrate the content and preparation of translational abstracts, the APA provides two practical examples from actual publications (APA, 2018). We utilized these established examples of translational abstracts in the redesign of the title pages from Study 1 and Study 2. Once again, we assessed the same experimental conditions (CB, CC, GB) as in the first two studies 3. We also used the same planned missing design and randomly assigned participants to the conditions using a balanced experimental plan.

Procedure and Statistical Analyses

The procedure was equivalent to the procedure followed in Studies 1 and 2. For a demo version of the survey with all three conditions, visit https://pub-demo.formr.org.

Instruments

We used the same instruments as in Studies 1 and 2 and tested factorial validity with the same series of (M)CFA models (see Table X). Again, internal consistencies were good for *integrity* ($\omega = .88$ and .90), acceptable for the four-item *topic specific multiplism* ($\omega = .69$ and .60) scale, and very good for the *treatment check* ($\omega = .84$ and .94).

Table 3

Results of the CFAs with fit indices (Study 3)

	1d CFA METI 1	1d CFA METI 2	3d CFA METI 1	3d CFA METI 2	1d MCFA METI	3d MCFA METI	1d CFA TSM 1	1d CFA TSM 2	1d MCFA TSM	1d CFA TCH 1	1d CFA TCH 2
χ^2	308.080	297.077	194.143	196.021	437.579	313.519	3.991	5.472	4.328	3.982	8.846
df	77.000	77.000	74.000	74.000	154.000	148.000	4.000	4.000	3.000	4.000	5.000
CFI	0.931	0.941	0.964	0.967	0.953	0.972	1.000	0.988	0.995	1.000	0.996
TLI	0.918	0.930	0.956	0.960	0.944	0.966	1.000	0.982	0.980	1.000	0.991
RMSEA	0.108	0.105	0.079	0.080	0.060	0.047	0.000	0.038	0.029	0.000	0.066
SRMR	0.040	0.033	0.031	0.025	0.159	0.089	0.030	0.050	0.092	0.012	0.016
SRMR between	-	-	-	-	0.052	0.040	-	-	0.071	-	-
SRMR within	-	-	-	-	0.107	0.049	-	-	0.020	-	-
BIC	9735.922	9077.757	9638.632	8993.349	18625.957	18539.351	2421.423	2341.021	4734.532	1709.467	1853.905
AIC	9636.548	8978.383	9528.610	8883.327	18329.002	18216.942	2385.932	2319.726	4645.445	1676.277	1822.088

Note: 1d: one-dimensional, 3d: three-dimensional, METI: Muenster Epistemic Trustworthiness Inventory, TSM: topic specific multiplism, TCH: treatment check

Sample

Participants were recruited in the UK general population via the online access panel provider *respondi*. Based on the latest UK census data (Office for National Statistics et al., 2016) we generated cross quotas of the variables sex, age, and qualification. In the survey, we used filter questions to achieve the same cross quota within our sample. By doing so, we exceeded the stopping rule from our preregistration by $n = 7$ participants, as cross quota cells only closed after the last participant from that

cell *finished* the survey, while further participants from that cell were still able to begin the survey until that point.

Results

Treatment check

Descriptively, the results of the treatment check (Figure 7) indicated that the participants read the explanations of the badges carefully and gave corresponding answers. Deviating from Study 1 and Study 2, participants more often assumed OSP in the control condition where no explicit information was given about data, code, and material sharing.

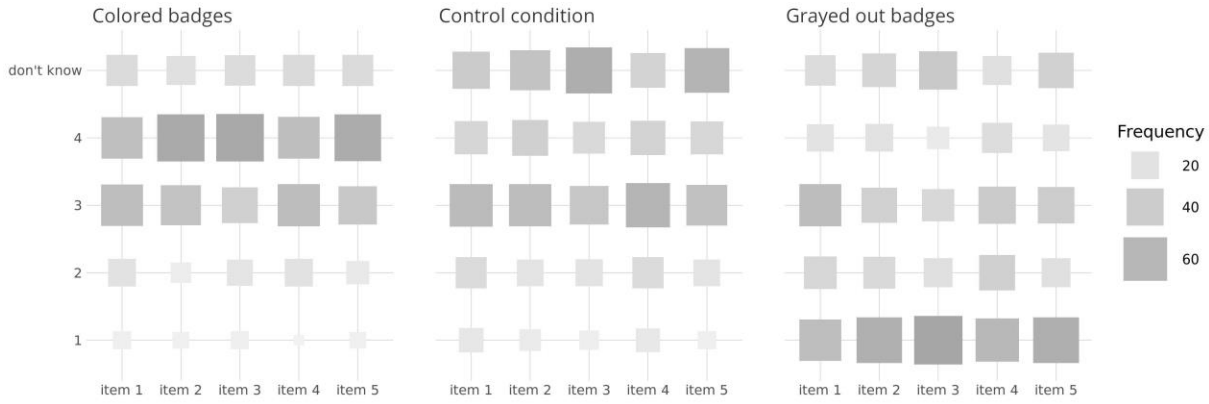

Figure 7

Fluctuation diagram of the results from the treatment check in Study 3. Frequency per item and experimental condition.

Hypothesis 1

Deviating from Studies 1 and 2, our data was more likely under $H1_2$ ($\mu(int)_{GB} < \mu(int)_{CC} = \mu(int)_{CB}$) than under $H1_1$ ($\mu(int)_{GB} = \mu(int)_{CC} = \mu(int)_{CB}$), which was reflected by the corresponding Bayes factors (BF against $H1_0$ = 3.2, BF against $H1_1$ = 5.8, BF against $\overline{H1}_2$ = 18.5, BF against $H1_u$ = 18.5). Nevertheless, participants of the public sample rated the integrity of researchers substantially lower

($d_{GB/CC}$ = .21, $d_{GB/CB}$ = .20) in the grayed out badges condition, but these ratings unexpectedly did not differ between the control condition and the condition with colored badges ($d_{CC/CB}$ = -.02).

Hypothesis 2

Regarding H2, we found strong evidence for the absence of an association between topic specific multiplism and integrity in all three conditions ($H2_0$: $b_1^{CB} = 0$ & $b_1^{CC} = 0$ & $b_1^{GB} = 0$; BF against $H2_1$ = 9.58, BF against $\overline{H2}_0$ = 33.5, BF against $H2_u$ = 33.5; see Figure 4).

(Exploratory) Hypothesis 3

Consistently, we found no evidence for the differences in associations between topic specific multiplism and integrity proposed by H3. Instead, the likelihood of the data was clearly greater for $H3_0$: $b_1^{GB} = b_1^{CC} = b_1^{CB} = 0$ in comparison to the alternatives stating an interaction (BF against $H3_1$ = 105.6, BF against $H3_2$ = 41.6, BF against $\overline{H3}_0$ = 31.2, BF against $H3_u$ = 31.2).

(Exploratory) Hypothesis 4

Finally, Study 3 also provided strong evidence for $H4_0$: $\mu(tsm)_{GB} = \mu(tsm)_{CC} = \mu(tsm)_{CB}$, meaning that the participants did, on average, report the same amount of topic specific multiplism in all three experimental conditions ($d_{GB/CC}$ = -.02, $d_{CC/CB}$ = -.13, $d_{GB/CB}$ = -.15; BF against $H4_1$ = 6.3, BF against $H4_2$ = 9.5, BF against $\overline{H4}_0$ = 22.4, BF against $H4_u$ = 22.4).

Discussion

Our findings substantiate the assumption that open science badges bear the considerable potential to influence trust in scientists as measured by perceived integrity. For undergraduates and scientists, we were able to corroborate the findings by Strømsø et al. (2011) on the negative relationship between multiplistic epistemic beliefs and epistemic trust. Moreover, we found evidence for the absence of a moderating effect of epistemic beliefs on the effects of badges on trust.

These results shed new light on the effects of badges. Beyond initial investigations of their effectiveness in fostering data sharing and adherence to open science standards (Kidwell et al., 2016),

we now have evidence that badges have the potential to increase trust in scientists by their target audiences (scientists and undergraduates). This is good news because knowing that higher trust is being given by the readership may strongly incentivize open science practices.

We argued that badges might be a tangible and contextualized way to signal adherence to standards compared to a simple communication strategy (e.g., Anvari & Lakens, 2018; Wingen et al., 2020). In the public sample, we were able to support this claim for visible noncompliance to OSP (GB condition), but not for visible compliance to OSP (CB condition). One explanation (also brought forward by Anvari and Lakens, 2018) may be that nonscientists believe that transparency is already fully ingrained in the scientific process. Our data is in line with this assumption. In fact, the treatment check revealed different perceptions of the researchers' OSP for the public sample versus undergraduates or scientists: Participants in the public sample more often assumed the adherence to OSP in the control condition compared to the two other samples. This potential "transparency assumption effect" still needs further investigation.

Our results should be qualified by the fact that we provided explanations of OSP in the texts that were situated in close proximity to the badges. These text-based specifications are also present in journals using badges (e.g., in *Psychological Science*), but in a less directly integrated format (e.g., at the end of the page). Research on different types of explanations or on alternatives to badges (e.g., using textual statements as in PLoS ONE) will give further insights into this matter.

Concerning multiplistic beliefs, our results are in line with previous research (Strømsø et al., 2011) on undergraduates and scientists. More specifically, the medium-sized negative effect of multiplism on perceived trustworthiness underpins the problematic nature of multiplistic beliefs in the context of information processing. As a side effect, utilizing badges to indicate that science is not just "opinion" triggered small decreases in topic-specific multiplistic beliefs. Important questions to clarify include determining the sustainability of these effects and whether they spill over onto domain-specific or general academic epistemic beliefs when individuals repeatedly perceive badges on publications badges (Author et al., 2018b).

In sum, our results further substantiate the assumption that badges produce desirable effects within their target audiences. This is good news for scientists and journal editors particularly because badges

are a simple and low-cost method of recognition (Kidwell et al., 2016). Nevertheless, it should be considered that the meaning and perception of badges are closely tied to the quality standards (and transparency) that decide when to award such a badge – an aspect that is also related to the question of who invests the resources to check the adherence to the standards and thus awards the badge. For example, a self-awarded badge on the researcher's personal website might not produce the same effects as badges awarded by journal editors who are guided by their transparent peer-review standards. Nevertheless, given the promising findings in our study, we conclude that OSP badges hold much potential, which is why we are excited about their further development and implementation.

References

Author et al. (2018a)

Author et al. (2018b)

Author et al. (2021)

Author et al. (in preparation)

Andrews Fearon, P., Götz, F. M., & Good, D. (2020). Pivotal moment for trust in science – don't waste it. *Nature*, *580*(7804), 456–456. https://doi.org/10.1038/d41586-020-01145-7

Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, *3*(3), 266–286. https://doi.org/10.1080/23743603.2019.1684822

APA. (2018, June). Guidance for Translational Abstracts and Public Significance Statements. Retrieved March 12, 2021, from American Psychological Association website: https://www.apa.org/pubs/journals/resources/translational-messages

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, *52*(1), 376–387. https://doi.org/10.3758/s13428-019-01236-y

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work?: Multiple imputation by chained equations. *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Bauer, P. J. (2020). Expanding the reach of psychological science. *Psychological Science*, *31*(1), 3–5. https://doi.org/10.1177/0956797619898664

Bråten, I., Britt, M. A., Strømsø, H. I., & Rouet, J.-F. (2011). The role of epistemic beliefs in the comprehension of multiple expository texts: Toward an integrated model. *Educational Psychologist*, *46*(1), 48–70. https://doi.org/10.1080/00461520.2011.538647

Bromme, R., & Goldman, S. R. (2014). The public's bounded understanding of science. *Educational Psychologist*, *49*(2), 59–69. https://doi.org/10.1080/00461520.2014.921572

Bromme, R., Kienhues, D., & Porsch, T. (2010). Who knows what and who can we believe? Epistemological beliefs are beliefs about knowledge (mostly) to be attained from others. In L. D.

Bendixen & F. C. Feucht (Eds.), *Personal epistemology in the classroom* (pp. 163–194). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511691904.006

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chang, M. K., Cheung, W., & Tang, M. (2013). Building trust online: Interactions among trust building mechanisms. *Information & Management*, *50*(7), 439–445. https://doi.org/10.1016/j.im.2013.06.003

Cochran-Smith, M. (2009). "Re-Culturing" teacher education: Inquiry, evidence, and action. *Journal of Teacher Education*, *60*(5), 458–468. https://doi.org/10.1177/0022487109347206

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Fecher, B., & Friesike, S. (2014). Open science: One term, five schools of thought. In S. Bartling & S. Friesike (Eds.), *Opening Science* (pp. 17–47). Cham: Springer International Publishing.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Franco, G. M., Muis, K. R., Kendeou, P., Ranellucci, J., Sampasivam, L., & Wang, X. (2012). Examining the influences of epistemic beliefs and knowledge representations on cognitive processing and conceptual change when learning physics. *Learning and Instruction*, *22*(1), 62–77. https://doi.org/10.1016/j.learninstruc.2011.06.003

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In I. B. Weiner (Ed.), *Handbook of Psychology* (pp. 87–114). Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/0471264385.wei0204

Grand, A., Wilkinson, C., Bultitude, K., & Winfield, A. F. T. (2012). Open science: A new "trust technology"? *Science Communication*, *34*(5), 679–689. https://doi.org/10.1177/1075547012443021

Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for Bayesian testing of order constrained hypotheses in structural equation models. *Journal of Statistical Computation and Simulation*, *89*(8), 1526–1553. https://doi.org/10.1080/00949655.2019.1590574

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261. https://doi.org/10.1111/bmsp.12110

Hendriks, F., & Kienhues, D. (2019). Science understanding between scientific literacy and trust: Contributions from psychological and educational research. In A. Leßmöllmann, M. Dascal, & T. Gloning (Eds.), *Science Communication* (pp. 29–50). Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110255522-002

Hendriks, F., Kienhues, D., & Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: The Muenster Epistemic Trustworthiness Inventory (METI). *PLOS ONE*, *10*(10). https://doi.org/10.1371/journal.pone.0139309

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research*, *67*(1), 88–140. https://doi.org/10.3102/00346543067001088

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Abingdon, Oxon: Chapman and Hall/CRC.

Hoijtink, H., Gu, X., Mulder, J., & Rosseel, Y. (2019). Computing Bayes factors from data with missing values. *Psychological Methods*, *24*(2), 253–268. https://doi.org/10.1037/met0000187

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*. https://doi.org/10.1037/met0000201

Isen, A. M. (2008). Some ways in which positive affect influences decisional making and problem solving. In *Handbook of Emotions* (pp. 548–573). New York: Guilford Press.

Kaslow, N. J. (2015). Translating psychological science to the public. *American Psychologist*, *70*(5), 361–371. https://doi.org/10.1037/a0039448

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., … Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective

method for increasing transparency. *PLoS Biology*, *14*(5), 1002456. https://doi.org/10.1371/journal.pbio.1002456

Kuhn, D., & Weinstock, M. (2002). What is epistemological thinking and why does it matter? In *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 121–144). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109–111. https://doi.org/10.1016/j.tics.2014.12.007

Laugksch, R. C. (2000). Scientific literacy: A conceptual overview. *Science Education*, *84*(1), 71–94. https://doi.org/10.1002/(SICI)1098-237X(200001)84:1<71::AID-SCE6>3.0.CO;2-C

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Liu, D., Vanderbilt, K. E., & Heyman, G. D. (2013). Selective trust: Children's use of intention and outcome of past testimony. *Developmental Psychology*, *49*(3), 439–445. https://doi.org/10.1037/a0031615

Lyon, L. (2016). Transparency: The emerging third dimension of open science and open data. *LIBER Quarterly*, *25*(4), 153–171. https://doi.org/10.18352/lq.10113

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734.

McCraw, B. W. (2015). The nature of epistemic trust. *Social Epistemology*, *29*(4), 413–430. https://doi.org/10.1080/02691728.2014.971907

Munthe, E., & Rogne, M. (2015). Research based teacher education. *Teaching and Teacher Education*, *46*, 17–24. https://doi.org/10.1016/j.tate.2014.10.006

Office for National Statistics, National Records of Scotland, & Northern Ireland Statistics and Research Agency. (2016). *2011 Census aggegate data* [Data set]. UK Data Service. https://doi.org/10.5257/CENSUS/AGGREGATE-2011-1

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Origgi, G. (2014). Epistemic trust. In P. Capet & T. Delavallade (Eds.), *Information evaluation* (pp. 35–54). Hoboken, NJ: John Wiley and Sons. https://doi.org/10.1002/9781118899151.ch2

Pew Research Center. (Ed.). (2019). *Trust and mistrust in Americans' views of scientific experts.* Retrieved from https://www.pewresearch.org/science/wp-content/uploads/sites/16/2019/08/PS_08.02.19_trust.in_.scientists_FULLREPORT_8.5.19.pdf

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Silvia, P. J., Kwapil, T. R., Walsh, M. A., & Myin-Germeys, I. (2014). Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior Research Methods*, *46*(1), 41–54. https://doi.org/10.3758/s13428-013-0353-y

Soderberg, C. K., Errington, T. M., & Nosek, B. A. (2020). *Credibility of preprints: An interdisciplinary survey of researchers* [Preprint]. MetaArXiv. https://doi.org/10.31222/osf.io/kabux

Stadtler, M., & Bromme, R. (2014). The content-source integration model: A taxonomic description of how readers comprehend conflicting scientific information. In *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 379–402). Cambridge, MA, US: MIT Press.

Stricker, J., Chasiotis, A., Kerwer, M., & Günther, A. (2020). Scientific abstracts and plain language summaries in psychology: A comparison based on readability indices. *PLOS ONE*, *15*(4), e0231160. https://doi.org/10.1371/journal.pone.0231160

Strømsø, H. I., Bråten, I., & Britt, M. A. (2011). Do students' beliefs about knowledge and knowing predict their judgement of texts' trustworthiness? *Educational Psychology*, *31*(2), 177–206. https://doi.org/10.1080/01443410.2010.538039

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton: CRC Press. Retrieved from https://stefvanbuuren.name/fimd/

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wickham, H., & Hofmann, H. (2011). Product plots. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2223–2230. https://doi.org/10.1109/TVCG.2011.227

Wingen, T., Berkessel, J. B., & Englich, B. (2020). No replication, no trust? How low replicability influences trust in psychology. *Social Psychological and Personality Science*, *11*(4), 454–463. https://doi.org/10.1177/1948550619877412

Zimmermann, M., & Jucks, R. (2018). How experts' use of medical technical jargon in different types of online health forums affects perceived information credibility: Randomized experiment with laypersons. *Journal of Medical Internet Research*, *20*(1), e30. https://doi.org/10.2196/jmir.8346