Research Article

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

Jürgen Schneider[1] , Tom Rosman[2] , Augustin Kelava[3] , and Samuel Merk[4]

[1]School of Education, University of Tübingen; [2]Department of Research Literacy and User Friendly Research Support, Leibniz Institute for Psychology Information; [3]Methods Center, University of Tübingen; and [4]Institute for School and Instructional Development in Primary and Secondary Education, University of Education Karlsruhe

Abstract

In three experimental studies, we investigated whether badges for open-science practices have the potential to affect trust in scientists and topic-specific epistemic beliefs by student teachers ($n = 270$), social scientists ($n = 250$), or the public ($n = 257$), all of whom were at least 16 years old. Furthermore, we analyzed the moderating role of epistemic beliefs for badges and trust. Each participant was randomly assigned to two of three conditions: badges awarded, badges not awarded, 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 with no badges) was not supported. For students and scientists, we found evidence for the relation of badges and epistemic beliefs as well as epistemic beliefs and trust. Further, we found evidence for the absence of moderation by epistemic beliefs.

In recent times, struggles to replicate empirical findings have been acknowledged in 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 awareness regarding open-science practices, a number of academic journals have adopted open-science badges. Because the authors' open-science badge disclosures are quality checked, readers can use badges to quickly determine whether a study has implemented open-science practices—an important indicator for gauging its transparency and trustworthiness. However, apart from some preliminary indications of their effectiveness in fostering the implementation of open-science practices

(Kidwell et al., 2016), not much is known about the effects of badges on perception at an individual level. Therefore, we investigated in three studies how trustworthy scientists are perceived to be by student teachers, 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

Corresponding Author:
Jürgen Schneider, University of Tübingen, School of Education
Email: juergen.schneider@uni-tuebingen.de

on other people's knowledge (Bromme et al., 2010). However, we cannot evaluate the truthfulness of all information from sources we interact with, particularly when we are 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) because 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 provide only overviews. Consequently, readers of scientific claims cannot evaluate the truthfulness of such claims by themselves but have to rely (to various degrees) on so-called secondhand evaluations (i.e., evaluations of the trustworthiness of an information source instead of the information itself; Bromme et al., 2010). Therefore, when people are acquiring and evaluating information, trust plays a pivotal role, as shown in studies on decision making (Isen, 2008; Liu et al., 2013) and learning (Landrum et al., 2015). This is equally true for different population groups, each interacting with scientific claims from their specific perspectives: scientists in their daily work, student teachers in their professional development (Munthe & Rogne, 2015), and the public through science communication (e.g., public health recommendations during a pandemic; Andrews Fearon et al., 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 et al., 1995) particularly highlight benevolence, integrity, and expertise as dimensions of trust (or, closely related, competence and warmth; Fiske et al., 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 who expose themselves to scrutiny by disclosing their scientific practices may help to rebuild trust in scientists (Grand et al., 2012), as it signals integrity on the part of the trusted researcher (Lyon, 2016). In line with these assumptions, a recent U.S. survey revealed that adults would trust scientific research findings more if the corresponding data were openly available (Funk et al.,

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 also in all disciplines involving empirical methods. Several journals currently use badges as a first indication of the authors' awareness of open-science practices. To our knowledge, our study is the first to investigate how these badges affect individual perceptions such as trust and epistemic beliefs. We found that badges increase trust in scientists and reduce multiplistic epistemic beliefs of student teachers and scientists. Given the current practice of authors self-reporting on their use of open-science practices, this result is a call to action concerning the accreditation process of badges. Furthermore, our results on epistemic beliefs indicate that badges may help to promote an idea of science that is not just an "opinion." Further results on the relation of trust and epistemic beliefs underscore their significance in information processing.

2019). Recently, these findings were corroborated for the German context by Rosman et al. (2022). Furthermore, Soderberg et al. (2020) reported 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 judgments.

In our view, badges are a tangible and contextualized way to signal adherence to or violation of standards concerning certain aspects of open-science practices (Bauer, 2020). Academic journals have increasingly adopted the practice of awarding open-science 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 open-science practices and a better adherence to open-science standards, particularly concerning data sharing (Kidwell et al., 2016).

Badges may suggest the certification of truthfully and responsibly implemented open-science practices, and we know from a study by Chang et al. (2013) that individual trust is particularly fostered by the mechanism of third-party certifications. However, practice in badge accreditation so far relies on authors to self-report which open-science practices they have implemented. With few exceptions, neither the journal nor the peer reviewers currently perform quality checks. Because we aimed in the present work to investigate
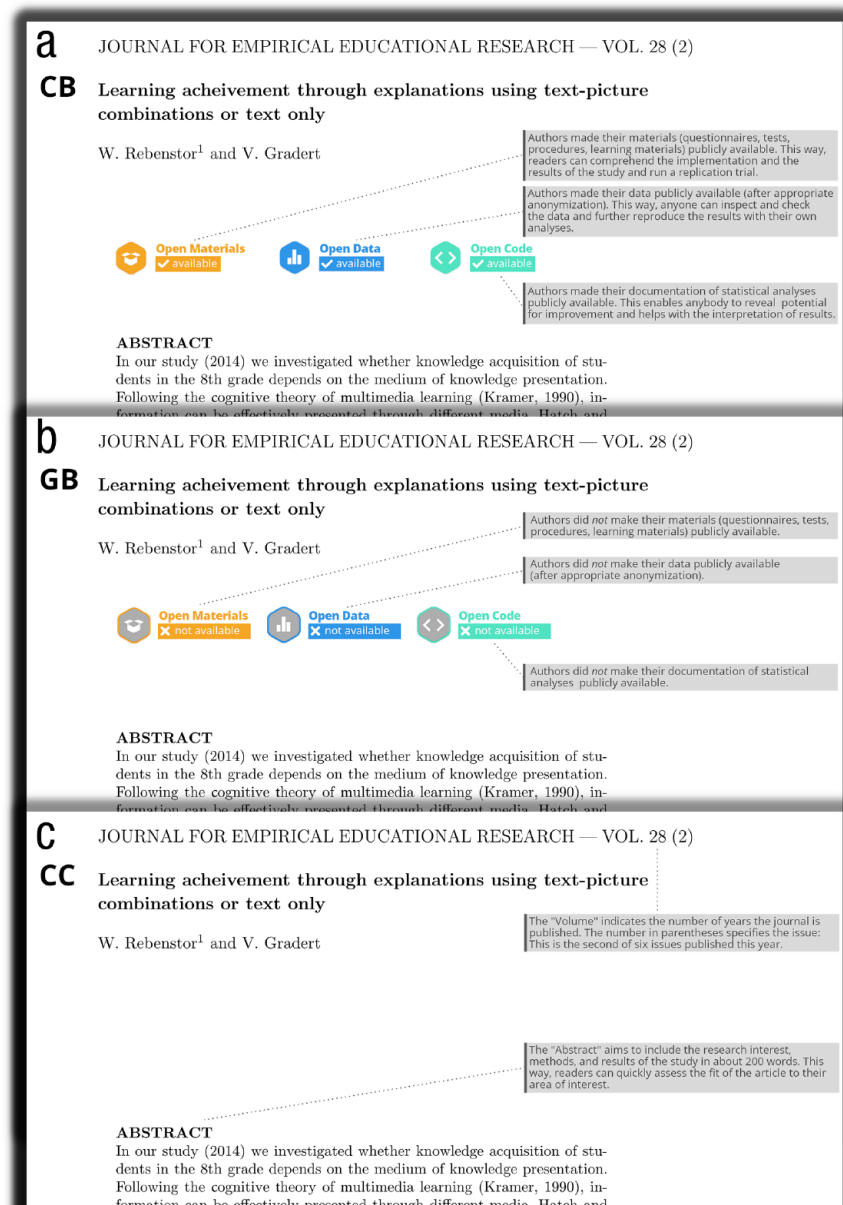
Fig. 1. Illustrations from the three experimental conditions: (a) colored badges (CB), (b) grayed-out badges (GB), and (c) control condition (CC). (Only the upper part of the title pages is shown here; for the full page, see the Supplemental Material available online.)

the potential effect of badges on trust, the badges had to be credibly associated with a responsible open-science practices implementation. In our study materials, we therefore vouched for a truthful and responsible open-science practices implementation by using additional explanations that explicitly describe how open-science practices were implemented in the fictitious studies (see Fig. 1).

We therefore argue that the badges displayed on scientists' publications or translational abstracts influence the perceived trustworthiness of the authors. Colored badges signal adherence to qualitative standards, increasing trust, and grayed-out badges signal violation of qualitative standards, decreasing trust compared with no badges. This prediction led to Hypothesis 1 (H1): Visible compliance to open-science practices (colored badges) leads to higher perceived trustworthiness of scientists compared with no information about open-science practices (control condition) or visible noncompliance to open-science practices (grayed-out badges), with visible noncompliance to open-science practices 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 people deal with textual information (Bråten et al., 2011; Franco et al., 2012). Developmental conceptualizations of epistemic beliefs distinguish between the consecutive stages of absolutism (knowledge as dualistic, i.e., 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 (Rosman et al., 2018) and negative relationships with judgments of text trustworthiness (Strømsø et al., 2011). This led to 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. This led to Hypothesis 3 (H3): Multiplistic epistemic beliefs moderate the effect of badges on perceived trustworthiness. However, because no corresponding empirical evidence exists to date, we label this hypothesis as exploratory.

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. This led to Hypothesis 4 (H4): Visible compliance to open-science practices (colored badges) leads to lower multiplistic epistemic beliefs compared with no information about open-science practices (control condition) or visible noncompliance with open-science practices (grayed-out badges). We concede, however, that this interpretation is somewhat speculative, which is why we again label this hypothesis as exploratory.

Study 1: Student Teachers

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 German teacher-education curricula (Cochran-Smith & Boston College Evidence Team, 2009).

Method

Given that we did not use methods that might have potentially harmed our participants (e.g., induction of fear, deception, or false feedback) and because all study data were collected anonymously, consistent with standard university procedure, no ethical clearance was sought. Before beginning the study, participants were required to sign an informed consent form informing them about their rights, data protection, and the study's methods and purposes.

Sample. Following our preregistration, we started recruiting the sample by advertising in social media groups and newsletters for student teachers from German universities. Our stopping rule was to cease data collection once a sample size of 270 was reached. We exceeded the stopping rule by 20 participants because the survey had to be deactivated manually after periodic sample-size checks. Thirteen participants skipped the repeated measurement, and four participants did not complete the demographic questions at the end of the questionnaire. On average, participants were 22.89 years old ($SD = 2.95$) and in their sixth semester ($M = 5.86$, $SD = 3.68$). Of all the participants, 176 indicated that they were female.

Design. Hypotheses were tested in an experiment with three conditions. Students were presented with two title pages of fictitious empirical journal articles (topics: dual-channel theory, learning by means of worked-out examples). These title pages contained either three colored badges with legends (colored-badges condition), three grayed-out badges with legends (grayed-out-badges condition), or no badges (control condition). These materials also included legends that explained other terms on the title page (see Fig. 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). Because we expected participants to be unfamiliar with badges, we included explanations of the badges in gray text boxes (see Fig. 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 open-science practices. To prevent experimental leakage, but at the same time increase test power, we used a planned-missing design (Graham et al.,

2003; Silvia et al., 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.

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 (TSM) scale (see below; Merk et al., 2018). Subsequently, they completed the Muenster Epistemic Trustworthiness Inventory (METI; Hendriks et al., 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 min 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 approximated adjusted fractional Bayes factors (BFs; Gu et al., 2018; Hoijtink, Mulder, et al., 2019) for informative hypotheses, as they are especially suitable to test hypotheses with order restrictions (Hoijtink, 2012) such as ours. To ensure a strictly confirmatory approach (Wagenmakers et al., 2012), we preregistered our hypotheses https://doi.org/10.17605/OSF.IO/YBS7F. Within this preregistration, we specified a data-analysis plan which served as a basis for our simulation-based sample-size determination (BF design analysis; see Schönbrodt & Wagenmakers, 2018). This data-analysis strategy and the results of the sample-size determination are described below.

BFs, 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 in order to gain the most compelling evidence. Our first hypothesis stated that student teachers would ascribe, on average, less integrity to the authors of studies if these title pages contained grayed-out badges compared with title pages that contained no information about the use of open-science practices, which, in turn, would be ascribed less integrity than authors of title pages containing colored badges. In our preregistration, we specified comparing this hypothesis $H1_1$: $\mu(\text{integrity})_{GB} < \mu(\text{integrity})_{CC} < \mu(\text{integrity})_{CB}$ with the corresponding point null hypothesis $H1_0$: $\mu(\text{integrity})_{GB} = \mu(\text{integrity})_{CC} =$ $\mu(\text{integrity})_{GB}$ and a hypothesis that assumes that only the visible adherence to open-science practices has an effect on integrity—$H1_2$: $\mu(\text{integrity})_{GB} = \mu(\text{integrity})_{CC} < \mu(\text{integrity})_{CB}$. (Note that throughout this article, the subscripts CB, GB, and CC refer to the colored-badges, grayed-out badges, and control conditions, respectively.) Furthermore, in our preregistration, we specified that if the data were to provide evidence for one of these hypotheses against the other two (Bayes factors: BF > 3 respective <1/3) and the corresponding hypothesis without constraints $H1_u$: $\mu(\text{integrity})_{GB}$; $\mu(\text{integrity})_{CC}$; $\mu(\text{integrity})_{CB}$, we would compare this hypothesis with 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 BFs outside the interval [1/3–3] would we consider our results as evidence for $H1_i$ and otherwise as inconclusive.

We computed these BFs using the routines implemented in the R package *bain* (Gu et al., 2019). This statistical package uses an adjusted and approximated version of the fractional BF, which 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 BFs using multiple-imputation data (Hoijtink, Gu, et al., 2019). We then imputed our (planned as well as unplanned) missing data using chained equations (Azur et al., 2011; van Buuren, 2012). Next, parameters of a repeated measures analysis of variance (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 a Cohen's *d* of .3 if $\mu(\text{integrity})_x \neq \mu(\text{integrity})_y$. The simulations suggested that a sample size of 250 would be sufficient, because 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).

Instruments. All studies were conducted using the Web-based survey tool formr (Arslan et al., 2020).

Integrity. The METI (Hendriks 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 interested only in one dimension of the inventory (see the 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 CFA, and its good model fit corroborated the assumption of three dimensions at the within-person level as well as at the between-person level (see Table 1 and the reproducible documentation of the analysis for details). Furthermore, all three-dimensional models significantly outperformed corresponding one-dimensional models (p values of χ^2 difference tests all $< .0001$). Because we specified τ-congeneric measurement models, McDonald's ω was used to assess internal consistency (Dunn et al., 2014), and this yielded good results, with a minimum score of ω = .83 (integrity in the first measurement).

Topic-specific multiplism. To assess TSM, we used a 4-point Likert-type scale by Merk et al. (2018; 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' open-science practices (using 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).

Results

Treatment check. Figure 2 depicts a fluctuation diagram (also known as a *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 with grayed-out badges and colored badges reveals that there were large effect sizes for ordinal measures (e.g., Varha & Delaney's A = .84 for Item 1). In the control condition, a high proportion of participants reported not knowing about the researchers' open-science practices, or their judgments showed high variation.

Hypothesis 1. H1 was that the colored-badges condition would induce higher perceived integrity of the authors than the control condition, which, in turn, would induce higher perceived integrity than the grayed-out-badges condition. To test H1, we applied the preregistered equation to compute the approximated adjusted fractional BFs for the corresponding Hypothesis $H1_1$: $\mu(\text{integrity})_{GB} < \mu(\text{integrity})_{CC} < \mu(\text{integrity})_{CB}$, the point null hypothesis $H1_0$: $\mu(\text{integrity})_{GB} = \mu(\text{integrity})_{CC} = \mu(\text{integrity})_{CB}$, and a hypothesis that postulated only an effect of the visible utilization on integrity, $H1_2$: $\mu(\text{integrity})_{GB} = \mu(\text{integrity})_{CC} < \mu(\text{integrity})_{CB}$, in which $\mu(\text{integrity})_X$ describes the mean of integrity in the group X (see the Statistical Analyses section). Because the underlying ANOVA model for such hypotheses assumes normality of the dependent variable, we first checked to see whether the data satisfied this assumption regarding skewness, kurtosis, and outliers. Because 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 $\underline{H1_0}$ = 3.5×10^7, 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 medium to large effect sizes, $d_{GB/CC}$ = 0.32, $d_{CC/CB}$ = 0.29, and $d_{GB/CB}$ = 0.57 (see Fig. 3).

Hypothesis 2. H2 predicted a negative association between TSM and integrity. To test this hypothesis, we specified a path model with three regression paths—one for each condition of TSM on integrity. Subsequently, we tested the hypothesis $H2_1$: $b_{1CB} > 0$ and $b_{1CC} > 0$ and $b_{1GB} > 0$ against $H2_0$: $b_{1CB} = 0$ and $b_{1CC} = 0$ and $b_{1GB} = 0$, again using the approximated adjusted fractional BF, which resulted in strong evidence for $H2_1$ (BF against $\underline{H2_0}$ = 6.0×10^{21}, BF against $\overline{H2_1}$ = 2.4×10^7, BF against $H2_u$ = 6.3). Table 2 depicts the pooled standardized regression coefficients as a measure of effect size.

Hypothesis 3 (exploratory). Table 2 also shows the results obtained for H3, which was that the association between TSM and integrity may be moderated by the condition, resulting in the following order of $H3_1$: $b_{1GB} > b_{1CC} > b_{1CB}$. We tested this hypothesis against the corresponding null hypothesis $H3_0$: $b_{1GB} = b_{1CC} = b_{1CB} = 0$ and a hypothesis that $H3_2$: $(b_{1GB}, b_{1CC}) > b_{1CB}$, meaning that

Table 1. Results and Fit Indices for the Confirmatory Factor Analyses (CFAs) in Study 1

Statistic	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	.997
TLI	.776	.715	.944	.950	.875	.971	.987	.985	.996	1.003	.991
RMSEA	.157	.208	.078	.087	.087	.042	.035	.045	.014	.000	.083
SRMR	.084	.099	.049	.040	.271	.172	.048	.055	.109	.006	.005
SRMR between subjects					.146	.091			.090	NA	NA
SRMR within subjects					.125	.081			.019	NA	NA
BIC	10,225.876	9,336.494	9,853.512	8,639.946	18,914.759	18,484.147	2,791.127	2,653.377	5,445.587	1,547.731	1,750.473
AIC	10,125.120	9,237.119	9,738.362	8,522.826	18,616.055	18,147.038	2,769.537	2,632.082	5,360.243	1,512.868	1,712.703

Note: METI = Muenster Epistemic Trustworthiness Inventory; MCFA = multilevel CFA; TSM = topic-specific multiplism; TCH = treatment check; CFI = comparative-fit index; TLI = Tucker–Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; BIC = Bayesian information criterion; AIC = Akaike information criterion.
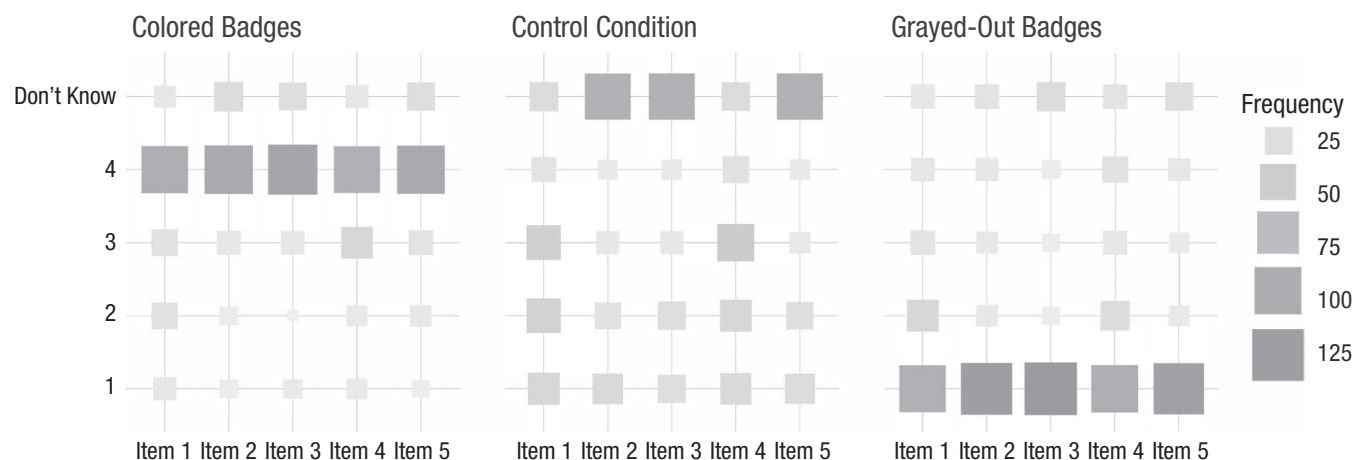
Fig. 2. Fluctuation diagram showing the frequency with which participants responded to each of the five items from the treatment check in Study 1. Results are shown separately for each experimental condition.

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 BFs 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 $H3_u = 18.5$).

Hypothesis 4 (exploratory). Finally, we tested whether the condition also had an effect on TSM. The violin plots depicted in Figure 4 indicate that there might be small to medium effects. This is underpinned by the effect-size estimates ($d_{GB/CC} = -0.26$, $d_{CC/CB} = 0.01$, $d_{GB/CB} = -0.25$) and the BFs that 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 show that TSM was smaller only when participants were

confronted with open-science-practices badges (BF against $H4_0 = 6.2$, BF against $H4_2 = 1.9$, BF against $\overline{H4}_1 = 8.4$, BF against $H4_u = 3.6$).

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 was expected to be more practiced in working with publications and might possibly have more knowledge of open-science badges.

Method

Sample. As the social sciences predominantly utilize empirical methods in research, we opted for a social-scientist sample. International participants were recruited
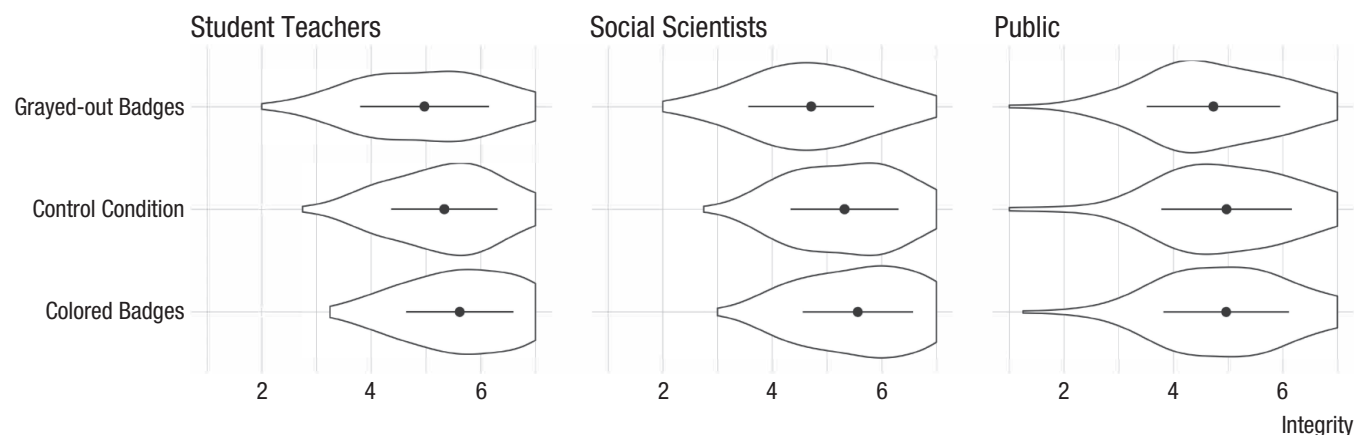

Fig. 3. Integrity ratings made by student teachers (Study 1), social scientists (Study 2), and the general public (Study 3) in each experimental condition. Violin plots show the density of the data, circles represent means, and error bars represent ±1 *SD*.

Table 2. Pooled Estimates (b_{1GB}, b_{1CC}, b_{1CB}) for All Three Studies for Integrity on Topic-Specific Multiplism in All Three Conditions

Study	Grayed-out badges (GB)	Control condition (CC)	Colored badges (CB)
Study 1 (student teachers)	−.44	−.41	−.32
Study 2 (social scientists)	−.35	−.37	−.41
Study 3 (general public)	−.03	−.13	−.09

via the online-access panel-provider prolific.co, filtering for social scientists with English as a first or fluent language. Following our stopping rule, we terminated data collection after 250 participants had passed the implemented quality check. No participant skipped the repeated measurement or demographic questions at the end of the questionnaire. Ninety-one participants were younger than 35 years old, 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 ($n = 91$); 170 participants identified as female.

Design. The design of the conditions was the same as in Study 1. To avoid potential bias in the participants' judgments because of topic familiarity (Tversky & Kahneman, 1973), we used abstracts of fictional studies (see the Supplemental Material available online). 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 three experimental conditions were utilized. 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 ($\omega = .91$ and $\omega = .92$), acceptable for TSM ($\omega = .69$ and $.64$), and very good for the treatment check ($\omega = .87$ and $.91$; see Table 3).

Results

Treatment check. As shown in Figure 5, Study 2 participants also complied very well with the treatment. The effect size for the first item comparing the colored-badges condition and the grayed-out-badges condition was even larger than in Study 1 (Varga & Delaney's A = .94).

Hypothesis 1. Figure 3 has already provided some insights with regard to $H1_1$: $\mu(\text{integrity})_{GB} < \mu(\text{integrity})_{CC} < \mu(\text{integrity})_{CB}$. Following the same (preregistered) procedure as in Study 1, we again obtained substantial relative evidence for $H1_1$ (BF against $H1_0 = 1.6 \times 10^{11}$, BF against $H1_2 = 7.5$, BF against $\overline{H1}_1 = 8.0 \times 10^2$, BF against
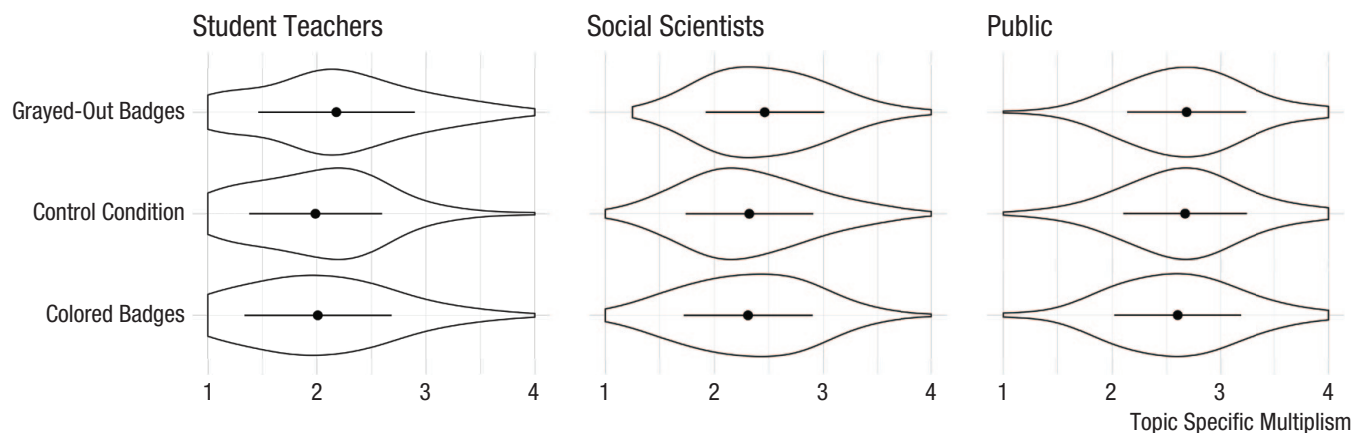

Fig. 4. Topic-specific multiplism ratings made by student teachers (Study 1), social scientists (Study 2), and the general public (Study 3) in each experimental condition. Violin plots show the density of the data, circles represent means, and error bars represent ±1 *SD*.
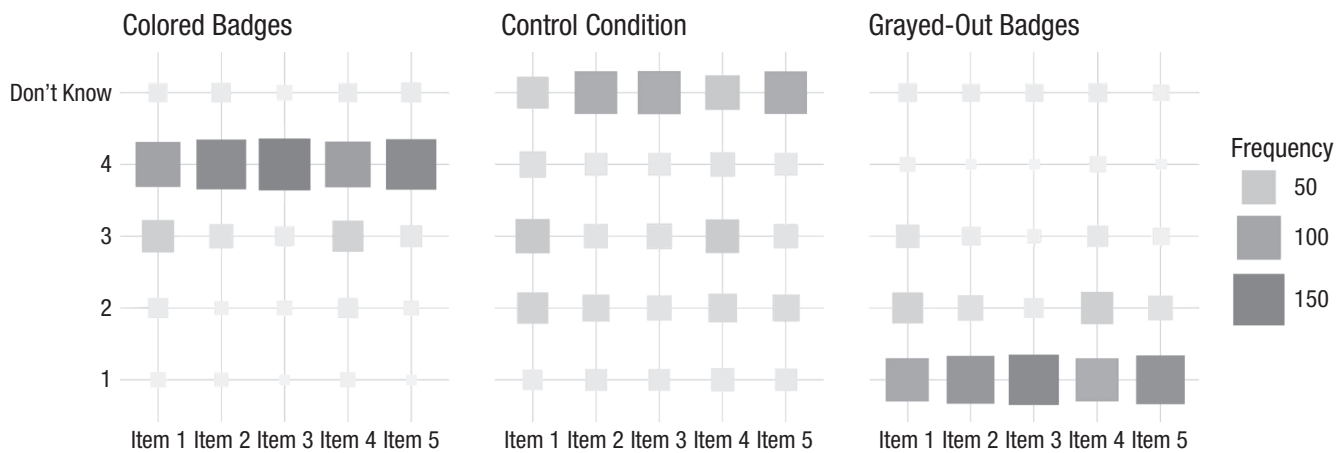
Fig. 5. Fluctuation diagram showing the frequency with which participants responded to each of the five items from the treatment check in Study 2. Results are shown separately for each experimental condition.

H1$_u$ = 5.4) with medium to large effect sizes, $d_{GB/CC}$ = 0.55, $d_{CC/CB}$ = 0.25, and $d_{GB/CB}$ = 0.77.

Hypothesis 2. In Study 2, the results regarding H2 were also replicated: Testing the hypothesis H2$_1$: $b_{1CB} > 0$ and $b_{1CC} > 0$ and $b_{1GB} > 0$ against H2$_0$: $b_{1CB} = 0$ and $b_{1CC} = 0$ and $b_{1GB} = 0$ revealed strong evidence for H2$_1$ (BF against H2$_0$ = 2.6 × 10^{16}, BF against $\overline{H2_1}$ = 2.6 × 10^7, BF against H2$_u$ = 6.9) with similar (medium) effect sizes to those in Study 1 (see Table 2).

Hypothesis 3 (exploratory). As in Study 1, the BFs found for H3 provided strong relative evidence for the null hypothesis (BF against $\underline{H3_1}$ = 1.3 × 10^2, BF against H3$_2$ = 1.2 × 10^2, BF against $\overline{H3_0}$ = 53.7, BF against H3$_u$ = 53.7).

Hypothesis 4 (exploratory). Finally, Study 2 revealed very similar results to Study 1 with regard to H4, showing moderately higher means in TSM for the condition with grayed-out badges ($d_{GB/CC}$ = −0.27, $d_{CC/CB}$ = 0.02, $d_{GB/CB}$ = −0.24), which is reflected by BFs 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

Sample. Participants were recruited from the UK general population via the online-access panel-provider respondi (https://www.respondi.com/EN/). Relying on the latest UK census data (Office for National Statistics et al., 2016), we generated cross quotas of three 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 seven participants, as cross-quota cells closed only after the last participant from that cell finished the survey; further participants from that cell were still able to begin the survey until that point.

Design. The experimental conditions were identical to those of 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 (Kerwer et al., 2021; Stricker et al., 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 is 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 (colored badges, control

Table 3. Results and Fit Indices for the Confirmatory Factor Analyses (CFAs) in Study 2

Statistic	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 subjects					0.232	0.085			0.019		
SRMR within subjects					0.072	0.040			0.008		
BIC	8,854.815	8,235.105	8,579.439	7,938.761	16,735.574	16,537.906	2,369.599	2,237.418	4,559.431	1,667.935	1,685.730
AIC	8,756.214	8,136.504	8,470.274	7,829.596	16,440.552	16,217.596	2,334.384	2,216.290	4,466.709	1,633.572	1,644.514

Note: METI = Muenster Epistemic Trustworthiness Inventory; MCFA = multilevel CFA; TSM = topic-specific multiplism; TCH = treatment check; CFI = comparative-fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; BIC = Bayesian information criterion; AIC = Akaike information criterion.

Fig. 6. Fluctuation diagram showing the frequency with which participants responded to each of the five items from the treatment check in Study 3. Results are shown separately for each experimental condition.

condition, grayed-out badges) as in the first two studies. 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 CFA and multilevel CFA (MCFA) models (see Table 4). Again, internal consistencies were good for integrity (ω = .88 and .90), acceptable for the four-item TSM (ω = .69 and .60) scale, and very good for the treatment check (ω = .84 and .94).

Results

Treatment check. Descriptively, the results of the treatment check (Fig. 6) indicated that the participants read the explanations of the badges carefully and gave corresponding answers. Deviating from Study 1 and Study 2, results showed that participants more often assumed that open-science practices were used in the control condition even though no explicit information was given there about data, code, and materials sharing.

Hypothesis 1. Deviating from Studies 1 and 2, our data was more likely under H1$_2$ (μ(integrity)$_{GB}$ < μ(integrity)$_{CC}$ = μ(integrity)$_{CB}$) than under H1$_1$ (μ(integrity)$_{GB}$ = μ(integrity)$_{CC}$ = μ(integrity)$_{CB}$), which was reflected in the corresponding BFs (BF against H1$_0$ = 3.2, BF against H1$_1$ = 5.8, BF against H1$_2$ = 18.5, BF against H1$_u$ = 18.5).

Nevertheless, participants of the public sample rated the integrity of researchers substantially lower ($d_{GB/CC}$ = 0.21, $d_{GB/CB}$ = 0.20) in the grayed-out-badges condition, but these ratings unexpectedly did not differ between the control condition and the condition with colored badges ($d_{CC/CB}$ = −0.02).

Hypothesis 2. Regarding H2, we found strong evidence for the absence of an association between TSM and integrity in all three conditions (H2$_0$: b$_{1CB}$ = 0 and b$_{1CC}$ = 0 and b$_{1GB}$ = 0; BF against H2$_1$ = 9.58, BF against H2$_0$ = 33.5, BF against H2$_u$ = 33.5; see Table 2).

Hypothesis 3 (exploratory). Consistently, we found no evidence for the differences in associations between TSM and integrity proposed by H3. Instead, the likelihood of the data was clearly greater for H3$_0$: b_{1GB} = b_{1CC} = b_{1CB} = 0 compared with the alternatives stating an interaction (BF against H3$_1$ = 105.6, BF against H3$_2$ = 41.6, BF against H3$_0$ = 31.2, BF against H3$_u$ = 31.2).

Hypothesis 4 (exploratory). Finally, Study 3 also provided strong evidence for H4$_0$: μ(TSM)$_{GB}$ = μ(TSM)$_{CC}$ = μ(TSM)$_{CB}$, meaning that the participants did, on average, report the same amount of TSM in all three experimental conditions ($d_{GB/CC}$ = −0.02, $d_{CC/CB}$ = −0.13, $d_{GB/CB}$ = −0.15; BF against H4$_1$ = 6.3, BF against H4$_2$ = 9.5, BF against H4$_0$ = 22.4, BF against H4$_u$ = 22.4).

Discussion

Our findings in two of the samples substantiate the assumption that open-science badges have considerable potential to influence trust in scientists as measured by perceived integrity as well as topic-specific

Table 4. Results and Fit Indices for the Confirmatory Factor Analyses (CFAs) in Study 3

Statistic	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 subjects					0.052	0.040			0.071		
SRMR within subjects					0.107	0.049			0.020		
BIC	9,735.922	9,077.757	9,638.632	8,993.349	18,625.957	18,539.351	2,421.423	2,341.021	4,734.532	1,709.467	1,853.905
AIC	9,636.548	8,978.383	9,528.610	8,883.327	18,329.002	18,216.942	2,385.932	2,319.726	4,645.445	1,676.277	1,822.088

Note: METI = Muenster Epistemic Trustworthiness Inventory; MCFA = multilevel CFA; TSM = topic-specific multiplism; TCH = treatment check; CFI = comparative-fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; BIC = Bayesian information criterion; AIC = Akaike information criterion.

multiplistic beliefs. For student teachers and scientists, we were able to corroborate findings 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 on perception. Beyond initial investigations of badges' 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 student teachers).

In the public sample, we were able to support this claim for visible noncompliance to open-science practices (grayed-out-badges condition), but not for visible compliance to open-science practices (colored-badges condition). One explanation (also proposed by Anvari & 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' open-science practices for the public sample versus the student teachers or scientists: Participants in the public sample more often assumed the adherence to open-science practices in the control condition compared with the two other samples. When the public has reason to believe that scientific practices are less transparent than they had assumed, perceived trust decreases accordingly. This potential "transparency assumption effect" still needs further investigation. Another question is the one regarding the evaluation of this potential effect: Should we avoid grayed-out badges in order to avoid decreasing trust in science? What speaks for avoiding grayed-out badges is that a lack of badges does not necessarily imply untrustworthy or low-quality scientific practice. Encouragingly, the participants in our public sample did not feel this way either: Even when trust decreased, trust scores remained in the upper range of the scale and did not turn into perceived untrustworthiness. However, it seems justified to assert that the public's perception changes on the basis of the transparency in research projects.

If badges are indeed related to trust, then this is an alarming call to action regarding the accreditation process. There is empirical evidence from the example of preregistrations that scientists often do not fully disclose deviations from their preregistrations (Claesen et al., 2021). Hence, badges may lull readers into a false sense of trust if they do not reflect reality. Should members of the public learn that the scientific community does not deliver what they expect of it, public trust in future research (Wingen et al., 2020) and past research

(Anvari & Lakens, 2018) could be severely affected beyond immediate repair. Therefore, if journals or websites adopt a badge system, it is crucial to implement third-party quality checks to ascertain that badges have been awarded for the right reasons.

Further, our studies contribute to research on epistemic beliefs. Our results are in line with previous research on the detrimental effect of multiplistic beliefs on trustworthiness among undergraduate students (Strømsø et al., 2011), and we expanded these findings to a sample of scientists. Particularly with respect to scientists, there have been few results on the correlates and structure of epistemic beliefs. More specifically, the medium to large negative effect of multiplism on perceived trustworthiness underpins the problematic nature of multiplistic beliefs in the context of information processing. Interestingly, these findings were not evident in the public sample. Could they, then, merely be an academic phenomenon? This would imply that epistemic-belief researchers might have to rethink whether epistemic beliefs are truly associated with trust or whether this is population specific. 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 (Merk et al., 2018). Also, concerning construct validity, we were able to further confirm the factor structure of epistemic beliefs with its subscales absolutism, multiplicism, and evaluativism, measured with Hendriks et al.'s (2015) survey instrument.

Our results should be qualified by the fact that we provided explanations of open-science practices 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). This might limit the generalizability of our findings; 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. Furthermore, in two of our studies, we recruited subjects from online panel providers. Evidence suggests that participants recruited in this manner may sometimes score lower on attention and honesty variables compared with traditional samples (Peer et al., 2021). In a comparative study by Peer et al. (2021), however, Prolific yielded the best scores in this regard. In addition, inattentive participants would have been filtered out by the attention checks used in our studies (Agley et al., 2022).

In sum, our results further substantiate the assumption that badges influence individual perceptions, particularly within their target audiences. This is good news under the assumption that open-science badges are "a simple, low-cost, effective method for increasing transparency" (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 determine whether 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 then awards the badge. Falsely awarded badges can turn the effect on trust into the opposite and cause lasting damage—a problem that can be solved only by rigorous quality control. Given the promising findings in our study, we conclude that badges on open-science practices hold much potential, which is why we are excited about their further development and implementation.

Transparency

ORCID iDs

Jürgen Schneider https://orcid.org/0000-0002-3772-4198
Tom Rosman https://orcid.org/0000-0002-5386-0499
Augustin Kelava https://orcid.org/0000-0001-6053-0415
Samuel Merk https://orcid.org/0000-0003-2594-5337

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976221097499

References

Agley, J., Xiao, Y., Nolan, R., & Golzarri-Arroyo, L. (2022). Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior Research Methods*, *54*(2), 885–897. https://doi.org/10.3758/s13428-021-01665-8

American Psychological Association. (2018, June). *Guidance for translational abstracts and public significance statements*. https://www.apa.org/pubs/journals/resources/translational-messages

Andrews Fearon, P., Götz, F. M., & Good, D. (2020). Pivotal moment for trust in science – Don't waste it. *Nature*, *580*(7804), 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

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? *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 University Press. https://doi.org/10.1017/CBO9780511691904.006

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L.,

Imai, T., . . . 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

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), Article 211037. https://doi.org/10.1098/rsos.211037

Cochran-Smith, M., & Boston College Evidence Team. (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). 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

Funk, C., Hefferon, M., Kennedy, B., & Johnson, C. (2019). *Trust and mistrust in Americans' views of scientific experts.* Pew Research Center. https://www.pewresearch.org/science/wp-content/uploads/sites/16/2019/08/PS_08.02.19_trust.in_.scientists_FULLREPORT_8.5.19.pdf

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). Wiley. 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). 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), Article e0139309. 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*. 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*, *24*(5), 539–556. https://doi.org/10.1037/met0000201

Isen, A. M. (2008). Some ways in which positive affect influences decisional making and problem solving. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 548–573). Guilford Press.

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

Kerwer, M., Chasiotis, A., Stricker, J., Günther, A., & Rosman, T. (2021). Straight from the scientist's mouth—plain language summaries promote laypeople's comprehension and knowledge acquisition when reading about individual research findings in psychology. *Collabra: Psychology*, *7*(1), Article 18898. https://doi.org/10.1525/collabra.18898

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14*(5), Article 1002456. https://doi.org/10.1371/journal.pbio.1002456

Kuhn, D., & Weinstock, M. (2002). What is epistemological thinking and why does it matter? In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 121–144). Erlbaum.

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

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

Merk, S., Rosman, T., Muis, K. R., Kelava, A., & Bohl, T. (2018). Topic specific epistemic beliefs: Extending the Theory of Integrated Domains in personal epistemology. *Learning and Instruction*, *56*, 84–97. https://doi.org/10.1016/j.learninstruc.2018.04.008

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, and 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), Article aac4716. https://doi.org/10.1126/science.aac4716

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

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-021-01694-3

Rosman, T., Bosnjak, M., Silber, H., Koßmann, J., & Heycke, T. (2022). Open science and public trust in science: Results from two studies. *Public Understanding of Science*. Advance online publication. https://doi.org/10.1177/09636625221100686.

Rosman, T., Peter, J., Mayer, A.-K., & Krampen, G. (2018). Conceptions of scientific knowledge influence learning of academic skills: Epistemic beliefs and the efficacy of information literacy instruction. *Studies in Higher Education*, *43*(1), 96–113. https://doi.org/10.1080/03075079.2016.1156666

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 D. N. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 379–402). 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), Article 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*. CRC Press. 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). 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), Article e30. https://doi.org/10.2196/jmir.8346