(Re)Building Trust? Journals' Open Science Badges Influence Trust in Scientists.

Running head

(Re)Building Trust?

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@uni-tuebingen.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 Tübingen, Department of School Education, Tübingen, Germany

Corresponding Author

Jürgen Schneider

juergen.schneider@uni-tuebingen.de

+49 7071 29 75331

Hausserstr. 43, 72074 Tübingen, Germany

(Re)Building Trust? Journals' Open Science Badges Influence Trust in Scientists.

Abstract

As a response to the replication crisis, reforms call for the implementation of open science standards. In this regard, open science badges are a promising method to signal a study's adherence to open science practices (OSP). In an experimental study, we investigated whether badges on journal article title pages affect non-scientists' trust in scientists. Furthermore, we analyzed the moderating role of epistemic beliefs in this regard. We randomly assigned 270 non-scientists to two of three conditions: Badges awarded (visible compliance to OSP), badges not awarded (visible non-compliance to OSP) and no badges (compliance not visible, control condition). Results indicate that badges influence trust in scientists as well as the epistemic beliefs of participants. However, epistemic beliefs did not moderate the effect of badges on trust. In sum, our paper provides support to the notion that badges are an effective means to promote epistemic beliefs and trust in scientists.

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. Badges are currently used by a number of journals in order to signal the compliance of individual articles to open science practices and to change the (so far adverse) incentive structures. 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 badges increase trust in scientists and reduce multiplistic epistemic beliefs. 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 put forward an idea of science as not just 'opinion'. Considering that badges are a low cost intervention, these results are encouraging.

Introduction

Recently, several scientific disciplines had to acknowledge their struggles in replicating empirical findings (Camerer et al., 2018; Open Science Collaboration, 2015). A primary reaction to this so-called replication crisis was the call for scientists to force transparency and reproducibility of the entire research process (Lindsay, 2015; Vazire, 2018). To signal the adherence to open science practices (OSP), a number of academic journals have adopted open science badges, which allow to quickly figure out whether the study implemented OSP - an important indicator to gauge its transparency and trustworthiness. However, beyond 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. In our study, we therefore investigate how trustworthy scientists are perceived depending on the inclusion of badges in their articles. Furthermore, considering the crucial role of beliefs about science in information processing, we explore the role of epistemic beliefs to moderate the effectiveness of badges and in directly predicting trust itself.

Epistemic trust

In our closely connected world that is characterized by division of cognitive labor, we are dependent on other people's knowledge (Bromme et al., 2010). However, we cannot evaluate the truthfulness of all information from people we interact with, particularly when lacking resources for its judgement such as knowledge, time and money (Stadtler & Bromme, 2014; Zimmermann & Jucks, 2018). When acquiring and evaluating information, trust plays a pivotal role, as has been shown in studies on decision-making (Isen, 2008; Liu et al., 2013) and learning (Landrum et al., 2015). This particularly applies to situations where non-scientists deal with socially relevant topics, such as public-health recommendations during a pandemic (Andrews Fearon et al., 2020), vaccines (Sharon et al., 2020), support for policies addressing climate change (Myers et al., 2017), or genetically modified organisms (Scanlon, 2020). We define non-scientists as individuals that have not received full formal education in scientific methods, and who therefore cannot evaluate the truthfulness of a scientific claim by themselves. Therefore, they have to rely, in most cases, 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). In many professional development programs that are not targeted toward a career in academia, students may nevertheless get in contact with scientific insights by reading scientific papers. This holds especially true for programs that aim toward professions like medical practitioners or teachers, in which evidence-based reflections play a central role (Cochran-Smith, 2009; Munthe & Rogne, 2015). For this population, trust in scientific knowledge may thus be particularly relevant.

On a conceptual level, we define trust as an attitude that describes the willingness of a trustor to make oneself vulnerable to actions of a trustee, where the trustor (implicitly or explicitly) forms predictions about the trustee's actions to be favorable toward him or her (McCraw, 2015). Research syntheses (e.g., Mayer et al., 1995) particularly highlight attributes of 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 with research reports on evidence generated by scientists.

Recent studies support the assumption of a detrimental effect of the replication crisis on perceived trustworthiness (Anvari & Lakens, 2019; Wingen et al., 2019). For this reason, the replication crisis was recently also framed as a crisis in credibility (Gall et al., 2017). OSP are increasingly called for and systematically encouraged as a response to challenges in replication and questionable research practices (European Commission, 2015; Lindsay, 2015). Alongside with goals of research quality and development (Fecher & Friesike, 2014), researchers exposing themselves to scrutiny by opening their scientific practices may help to rebuild trust in scientists (Grand et al., 2012) as it signals integrity on the side of the (trusted) researcher (Lyon, 2016). In line with these assumptions, a recent survey shows that U.S. adults state they would trust scientific research findings more, if the corresponding data were openly available (Pew Research Center, 2019).

Based on such findings, one may tentatively conclude that the apparent answer to a credibility crisis would be to inform non-scientists about the scientists' commitment to open reforms. However, recent research suggests that a straightforward communication strategy is not enough to rebuild trust in past research (Wingen et al., 2019), and that it may even further decrease trust in future research (Anvari & Lakens, 2019). The interventions used in these studies implemented rather decontextualized descriptions of OSP on a discipline-specific level. Participants not familiar with the scientific research process (i.e. non-scientists) might therefore not fully comprehend how these rather abstract 'reforms' shape research practice and why they consequently might help to improve replicability.

In our view, badges are a more tangible and contextualized way to signal the adherence to or violation of standards concerning a variety of OSP (Bauer, 2020). Over the last years, badges were increasingly adopted among academic journals (for a list, see https://www.cos.io/our-services/badges). First investigations indicate that badges are related to higher frequency of OSP and better adherence to its standards, particularly concerning data sharing (Kidwell et al., 2016). We therefore argue that badges on scientists' publications influence perceived trustworthiness of the authors, with badges signaling the adherence to standards increasing trust and badges signaling the violation of standards decreasing trust compared to no badges.

Hypothesis 1 (H^1): Visible OSP (badges) lead to higher perceived trustworthiness of scientists compared to no information about OSP or visible rejection of OSP (greyed out badges), with visible rejection of OSP receiving lowest trustworthiness ratings.

Epistemic beliefs and epistemic trust

Epistemic beliefs - individual perceptions about the nature of knowledge and knowing (Hofer & Pintrich, 1997)- are known to influence information processing when dealing 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, 'right-or-wrong'), multiplism and evaluativism (knowledge as weighed evidence). Because of their focus on personal opinions over facts and evidence, particularly multiplistic beliefs (knowledge as subjective opinions; Kuhn & Weinstock, 2002) seem to impair information processing - as evidenced by their negative effects on learning processes (Rosman et al., 2018), and negative relationships with judgements of text trustworthiness (Strømsø et al., 2011).

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

Furthermore, multiplistic beliefs depict the source of knowledge to be internalized in a knowing subject in forms of individual opinions. Individuals with high multiplistic beliefs thus see external sources of knowledge (e.g., researchers) and knowledge evaluation (e.g., badges) as irrelevant since they consider all knowledge claims to be equally true (Kuhn & Weinstock, 2002). For individuals with high multiplistic beliefs, the question 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 multiplistic beliefs, badges will not play a role regarding their epistemic trust. Since no corresponding empirical evidence exists to date, we, however, labeled this hypothesis as exploratory.

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

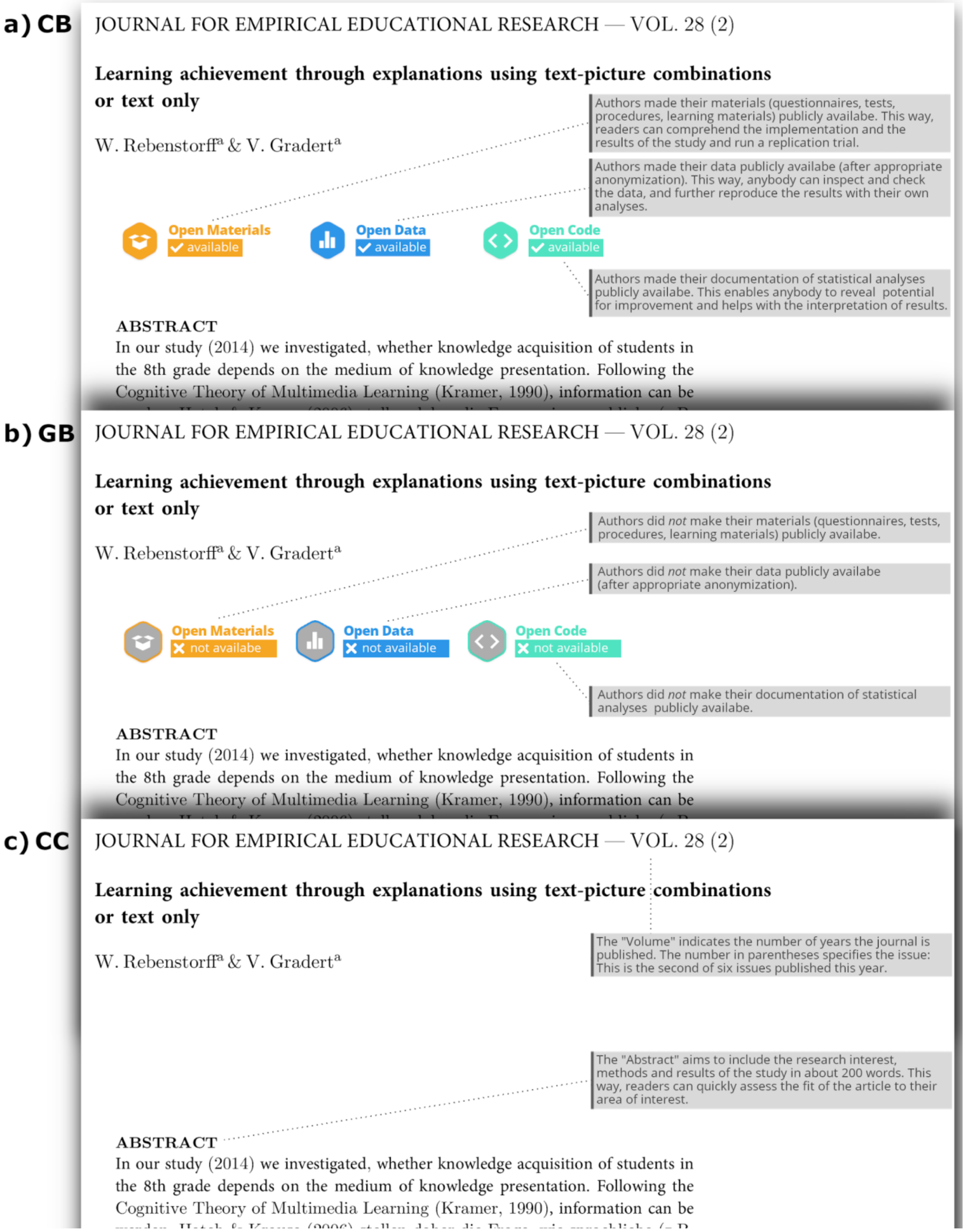
Moreover, badges might indicate that science is not just ‘opinion’ since they make the underlying empirical and fact-based approach more tangible, thus leading to reductions in multiplistic beliefs. We however concede that this interpretation is somewhat speculative, which is why we, again, label the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

Hypothesis 4 (H^4): Visible OSP (badges) lead to lower multiplistic epistemic beliefs compared to no information about OSP or visible rejection of OSP (greyed out badges).

Method

Design

Hypotheses were tested in an experiment with three conditions in the context of teacher education: Students were presented two title pages of fictitious empirical journal articles (topics: dual channel theory, learning with worked out examples) whereby these title pages contained either a) three colored badges with legends (condition “colored badges”, CB), or b) three greyed out badges with legends (condition ”greyed out badges”, GB), or c) no badges (“control condition”, CC), but also legends which explained other terms of the title page (see Fig. 1). The three colored badges indicated that the



JOURNAL FOR EMPIRICAL EDUCATIONAL RESEARCH — VOL. 28 (2)

Learning achievement through explanations using text-picture combinations or text only

W. Rebenstorff[a] & V. Gradert[a]

Authors made their materials (questionnaires, tests, procedures, learning materials) publicly availabe. This way, readers can comprehend the implementation and the results of the study and run a replication trial.

Authors made their data publicly availabe (after appropriate anonymization). This way, anybody can inspect and check the data, and further reproduce the results with their own analyses.

Authors made their documentation of statistical analyses publicly availabe. This enables anybody to reveal potential for improvement and helps with the interpretation of results.

ABSTRACT

In our study (2014) we investigated, whether knowledge acquisition of students in the 8th grade depends on the medium of knowledge presentation. Following the Cognitive Theory of Multimedia Learning (Kramer, 1990), information can be

JOURNAL FOR EMPIRICAL EDUCATIONAL RESEARCH — VOL. 28 (2)

Learning achievement through explanations using text-picture combinations or text only

W. Rebenstorff[a] & V. Gradert[a]

Authors did *not* make their materials (questionnaires, tests, procedures, learning materials) publicly availabe.

Authors did *not* make their data publicly availabe (after appropriate anonymization).

Authors did *not* make their documentation of statistical analyses publicly availabe.

ABSTRACT

In our study (2014) we investigated, whether knowledge acquisition of students in the 8th grade depends on the medium of knowledge presentation. Following the Cognitive Theory of Multimedia Learning (Kramer, 1990), information can be

JOURNAL FOR EMPIRICAL EDUCATIONAL RESEARCH — VOL. 28 (2)

Learning achievement through explanations using text-picture combinations or text only

W. Rebenstorff[a] & V. Gradert[a]

The "Volume" indicates the number of years the journal is published. The number in parentheses specifies the issue: This is the second of six issues published this year.

The "Abstract" aims to include the research interest, methods and results of the study in about 200 words. This way, readers can quickly assess the fit of the article to their area of interest.

ABSTRACT

In our study (2014) we investigated, whether knowledge acquisition of students in the 8th grade depends on the medium of knowledge presentation. Following the Cognitive Theory of Multimedia Learning (Kramer, 1990), information can be

Fig. 1.

Illustrations of the three experimental conditions (upper part of the title pages). a) CC: Colored badges, b) GB: greyed out badges, c) CC: control condition authors implemented the open science practices “open data”, “open materials” and “open code”, whereas greyed out badges signaled non-adherence with these practices “data not available”, “materials not available” and “code not available”. As we expected participants not to be familiar with badges, we integrated explanations of the badges in grey text boxes. These were explicitly labelled as additional explanations, that are not part of the journal article itself (see Fig. 1). In the condition without badges, participants did not receive information about the implementation of OSPs. 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): Every participant went through two of the three conditions. The assignment and sequence of these conditions as well as the topics and the sequence of the topics were randomized using a balanced experimental plan.

Procedure

After participants agreed to the informed consent, they were introduced into the procedure of the survey and informed about its structure. They were told that they’ll be given the title page of a regular journal article with explanations annotated in grey text boxes. Participants were asked to thoroughly read the title page and answer the questions below. On the next survey page participants read the first title page and were prompted to respond to a topic-specific multiplism scale below (Merk et al., 2018). Subsequently, they gave their answers on the Muenster Epistemic Trustworthy Inventory (METI, Hendriks et al., 2015), and finally processed the treatment check. This sequence of events was repeated for the second title page. Finally, at the end of the questionnaire, participants were subjected to a few demographic questions. The survey took approximately fifteen minutes to complete, for a demo version of the survey see (removed for blind review).

Statistical Analyses

We planned to analyze our data using (approximate adjusted fractional) Bayes factors (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), like ours. To ensure a strictly confirmatory approach (Wagenmakers et al., 2012), we preregistered our hypotheses (removed for blind 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 how more likely the current data are to be observed under a specific hypothesis, in contrast to another hypothesis. In this regard, a central challenge is choosing which hypotheses to compare in order to gain the most compelling evidence. Consider Hypothesis 1 (H^1), were we stated that student teachers on average would ascribe less integrity (int) to the authors of studies if these title pages contained greyed 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). We preregistered to compare this hypothesis H_1^1: $\mu(\text{int})_{\text{GB}} < \mu(\text{int})_{\text{CC}} < \mu(\text{int})_{\text{CB}}$ with the corresponding point null-hypothesis H_0^1: $\mu(\text{int})_{\text{GB}} = \mu(\text{int})_{\text{CC}} = \mu(\text{int})_{\text{CB}}$ and a hypothesis that assumes that only the visible utilization of OSP has an effect on integrity (H_2^1: $\mu(\text{int})_{\text{GB}} = \mu(\text{int})_{\text{CC}} < \mu(\text{int})_{\text{CB}}$. Furthermore, we preregistered that if the data provides evidence for one of these hypotheses against the two others (BF > 3 respective < 1/3) and the corresponding hypothesis without constraints H_u^1: $\mu(\text{int})_{\text{GB}}$; $\mu(\text{int})_{\text{CC}}$; $\mu(\text{int})_{\text{CB}}$, we would compare this hypothesis to its complement $\overline{H_i^1}$ (which contains all mean configuration which don't satisfy the restrictions of H_i^1) and only if all these comparisons also resulted in Bayes factors outside the interval [1/3; 3], we would judge our results as evidence for H_i and otherwise as inconclusive.

We computed these Bayes factors using the routines implemented in the R package bain (Gu et al., 2019). It 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 purpose as routines for computing Bayes factors using multiple imputation data exist (Hoijtink, Gu, et al., 2019). Correspondingly, we imputed our (planned as well as unplanned) missing data using chained equations (Azur et al., 2011; van Buuren, 2018). Thereupon, parameters of a repeated measurement ANOVA were estimated on each of the resulting (1000) complete data sets and combined using the rules derived by Hoijtink et al. (2019).

To determine our preregistered sample size, we ran simulation studies which used the decision procedure described above and assumed Cohen's d = .3 if $\mu(\text{int})_x \neq \mu(\text{int})_y$. The simulations suggested that a sample size of 250 would be sufficient, as, in the worst case (true hypothesis is H_2^1), our decision procedure would result in evidence for a wrong 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 German universities. According to our stopping rule, we stopped data collection at N = 270. Thirteen participants skipped the repeated measurement and four skipped 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), and 176 were of female gender.

Instruments

Integrity.

We used the Muenster Epistemic Trustworthiness Inventory (METI, Hendriks et al., 2015) to assess the amount of integrity participants ascribed towards the authors of the respective title page. This instrument uses 14 antonym-pairs which have to be rated on a 7-point scale and are mapped to three subscales (expertise: *well educated - poorly educated*; integrity: *honest - dishonest*; benevolence: *considerate - inconsiderate*). Despite the fact that we were only interested in one dimension of the inventory (see preregistration), participants were provided all three dimensions since we wanted to additionally gain some insights about the construct validity of the instrument and have more covariates to multiply impute the planned missing data. Therefore, we first carried out a confirmatory factor analyses (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, whose good model fit corroborated the assumption of three dimensions at the within-person as well as at between-person levels (see Table 1 and the reproducible documentation of the analysis [RDA] for details). Furthermore, all three dimensional models outperformed corresponding one dimensional models significantly (*p*-values of χ^2-difference test all smaller than .0001). As we specified τ-congeneric measurement models, McDonald's ω was used to assess internal consistency (Dunn et al., 2014) which 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 Merk, et al, (2018, sample item: "*The insights from the text are arbitrary*"). Consecutive as well as two-level CFA 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 .53).

Treatment check.

In terms of compliance to our treatment, we examined, if participants recognized and understood the (partially greyed out) badges. To investigate their compliance, we directly and indirectly asked participants 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 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 .90).

Table 1.

Results of the CFAs with fit indices

	1d	**1d**	**3d**	**3d**	**1d**	**3d**	**1d**	**1d**	**1d**	**1d**	**1d**
	CFA	**CFA**	**CFA**	**CFA**	**MCFA**	**MCFA**	**CFA**	**CFA**	**MCFA**	**CFA**	**CFA**
	METI 1	**METI 2**	**METI 1**	**METI 2**	**METI**	**METI**	**TSM 1**	**TSM 2**	**TSM**	**TCH 1**	**TCH 2**
χ^2	588.932	936.555	194.174	212.262	764.777	277.759	5.302	6.072	4.422	.359	.505
df	77.000	77.000	73.000	72.000	154.000	145.000	4.000	4.000	4.000	1.000	1.000
CFI	.811	.759	.955	.961	.894	.977	.991	.990	.999	1.000	1.000
TLI	.776	.715	.944	.950	.875	.971	.987	.985	.996	1.004	1.011
RMSEA	.157	.208	.078	.087	.087	.042	.035	.045	.014	.000	.000
SRMR	.084	.099	.049	.040	.271	.172	.048	.055	.109	.002	.005
SRMR between	-	-	-	-	.146	.091	-	-	.090	-	-
SRMR within	-	-	-	-	.125	.081	-	-	.019	-	-
BIC	10225.876	9336.494	9853.512	8639.946	18914.759	18484.147	2791.127	2653.377	5445.587	1931.815	1061.585
AIC	10125.120	9237.119	9738.362	8522.826	18616.055	18147.038	2769.537	2632.082	5360.243	1901.521	1037.363

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

Results

Treatment check

Figure 2 shows a fluctuation diagram (also known as “product plot“, Wickham & Hofmann, 2011) of the treatment check’s results. We judge these results as evidence for a strong compliance with our treatment, as, for example, comparing the condition GB and CB resulted in large effect sizes for ordinal measures (e.g., Varha & Delaney’s A = .12 for Item 1). In the CC, participants increasingly reported not to know about the researchers’ OSPs or the judgements showed high variation.

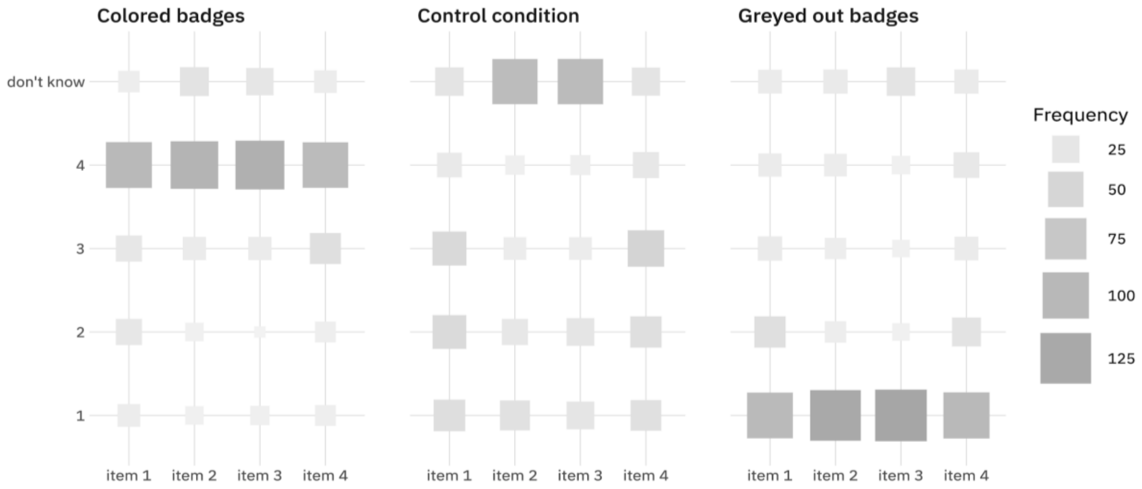


Fig. 2.

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

Hypothesis 1

Hypothesis 1 (H^1) 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 H^1, we preregistered to compute approximate adjusted fractional Bayes factors for the corresponding Hypothesis H_1^1: $\mu(\text{int})_{GB} < \mu(\text{int})_{CC} < \mu(\text{int})_{CB}$, the point null-hypothesis H_0^1: $\mu(\text{int})_{GB} = \mu(\text{int})_{CC} = \mu(\text{int})_{CB}$, and a hypothesis that postulates only an effect of the visible utilization on integrity H_2^1: $\mu(\text{int})_{GB} = \mu(\text{int})_{CC} < \mu(\text{int})_{CB}$, whereby $\mu(\text{int})_X$ describes the mean of integrity in the group X (see section Statistical analysis). As the underlying ANOVA model for such hypotheses assumes normality of the dependent variable, we first checked if the data satisfied this assumption in terms of skewness, kurtosis and outliers. As the data showed no strong violations of these criteria, we went on by (multiply) imputing the planned and unplanned missing data using the procedures implemented in the mice package (Van Buuren & Groothuis-Oudshoorn, 2011). Using this data, we followed the preregistered decision procedures described in the "Statistical analyses" section. This resulted in substantial evidence for H_1^1 as all Bayes factors exceeded 5.5 in favor to H_u^1 (BF against $H_0^1 = 3.5 \cdot 10^7$, BF against $H_2^1 = 4.5 \cdot 10^1$, BF against $\overline{H_1^1} = 4.8 \cdot 10^3$, BF against $H_u^1 = 5.5$).

Furthermore, comparing the means of integrity between the three experimental groups resulted in moderate to large effect sizes $A/d_{GB/CC}$ = .41/-.32, $A/d_{CC/CB}$ = .42/-.29, $A/d_{GB/CB}$ = .34/-.57 whereby *A* denotes Vargha and Delaney's A which is a measure of stochastic superiority and identical to the common language effect size (McGraw & Wong, 1992) if the measure is continuous.

Hypothesis 2

Hypothesis 2 (H^2) 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 from topic specific multiplism on integrity (see Figure 3). Subsequently, we tested the hypothesis H_1^2: $b_1^{CB} > 0$ & $b_1^{CC} > 0$ & $b_1^{GB} > 0$ against H_0^2: $b_1^{CB} = 0$ & $b_1^{CC} = 0$ & $b_1^{GB} = 0$, again using the approximate adjusted fractional Bayes factor which resulted in strong evidence for H_1^2 (BF against $H_0^2 = 6.0 \cdot 10^{21}$, BF against $\overline{H_1^2} = 2.4 \cdot 10^7$, BF against $H_u^2 = 6.3$). Figure 3 depicts the pooled standardized regression coefficients as a measure of effect size.

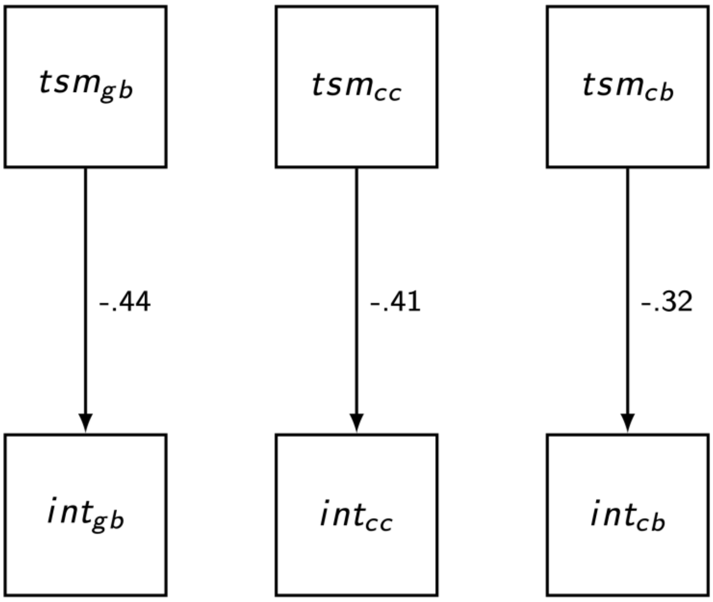


Fig. 3.

Path model for hypotheses 3 and 4 with pooled estimates. For clarity we did not depict variances and covariances. tsm = topic specific multiplism, int = integrity, gb = greyed out badges, cc = control condition, cb = colored badges.

(Exploratory) Hypothesis 3

From Figure 3 one can obtain information about the results regarding Hypothesis 3 which states, that the association between topic specific multiplism and integrity may be moderated by the topic resulting in the following order of H_1^3: $\mathrm{b_1}^{\mathrm{GB}} > \mathrm{b_1}^{\mathrm{CC}} > \mathrm{b_1}^{\mathrm{CB}}$. We tested this hypothesis against the corresponding null-hypothesis H_0^3: $\mathrm{b_1}^{\mathrm{GB}} = \mathrm{b_1}^{\mathrm{CC}} = \mathrm{b_1}^{\mathrm{CB}} = 0$ and a hypothesis which states H_2^3: $(\mathrm{b_1}^{\mathrm{GB}}, \mathrm{b_1}^{\mathrm{CC}}) > \mathrm{b_1}^{\mathrm{CB}}$ which means that the association is smaller when participants saw that scientists used 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 $\mathrm{H}_1^3 = 6.0$, BF against $\mathrm{H}_2^3 = 7.4$, BF against $\overline{\mathrm{H}_0^3} = 18.5$, BF against $\mathrm{H}_u^3 = 18.5$).

(Exploratory) Hypothesis 4

Finally, we tested if the condition also had an effect on topic specific multiplism. Figure 4 already indicates that there might be small to moderate effects. This is underpinned by the effect size estimates ($A/d_{GB/CC} = .41/-.32$, $A/d_{CC/CB} = .42/-.29$, $A/d_{GB/CB} = .34/-.57$) and the Bayes factors which favors H_1^4: $\mu(\mathrm{tsm})_{\mathrm{GB}} > \mu(\mathrm{tsm})_{\mathrm{CC}} > \mu(\mathrm{tsm})_{\mathrm{CB}}$ against a corresponding null-hypothesis H_0^4: $\mu(\mathrm{tsm})_{\mathrm{GB}} = \mu(\mathrm{tsm})_{\mathrm{CC}} = \mu(\mathrm{tsm})_{\mathrm{CB}}$ and a less specific hypothesis H_2^4: $(\mu(\mathrm{tsm})_{\mathrm{GB}}, \mu(\mathrm{tsm})_{\mathrm{CC}}) > \mu(\mathrm{tsm})_{\mathrm{CB}}$ which states only, that topic specific multiplism is smaller when participants are confronted with OSP badges (BF against $\mathrm{H}_0^4 = 6.3$, BF against $\mathrm{H}_2^4 = 3.3$, BF against $\overline{\mathrm{H}_1^4} = 8.5$, BF against $\mathrm{H}_u^4 = 3.5$).

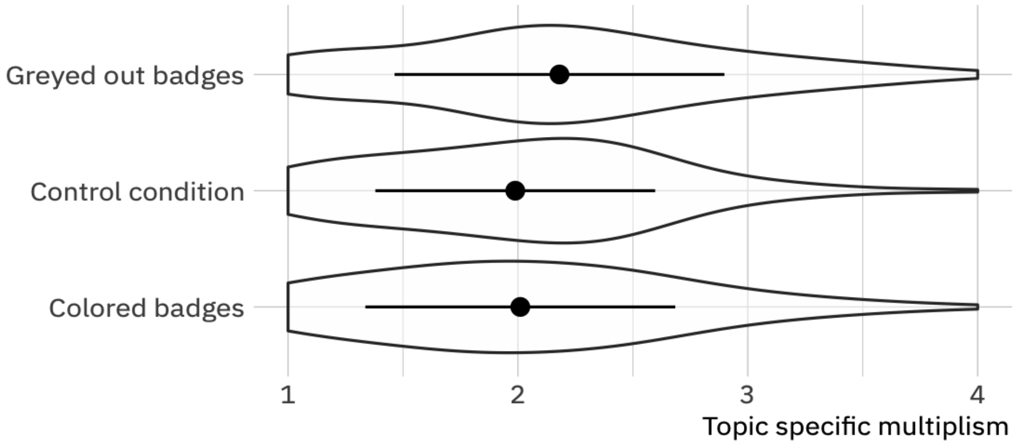


Fig. 4.

Graphical overview of H^4. Violin plots and means ±1*SD.

Discussion

Our findings substantiate the assumption that open science badges bear considerable potential to influence trust in scientists as measured by perceived integrity, as well as epistemic beliefs related to the topic in question. Moreover, we were able to further support findings by Strømsø et al. (2011) on the negative relationship of multiplistic epistemic beliefs and epistemic trust. Multiplistic epistemic beliefs, however, do not seem to moderate the effects of badges on trust.

These results shed new light on the effects of badges. Beyond first investigations on their effectiveness in fostering data sharing and adherence to open science standards (Kidwell, 2016), we now know that badges are also suited to increase trust in scientists. This is good news since higher trust awarded from non-scientists may even further incentivize open science practices. A significant question for further research is whether scientists share the same feelings about their colleagues receiving badges.

We argued that badges might be a tangible and contextualized way to signal the adherence to standards as compared to simple communication strategy (e.g., Wingen et al., 2019; Anvari & Lakens, 2019). One explanation by Anvari and Lakens (2019) on the ineffectiveness of the communication strategy was that non-scientists might believe transparency to already be ingrained in the scientific process. Our data did not support this assumption as participants in the condition without badges and participants in the condition with badges indicating the adherence to OSP differed in their perceptions of the researchers' OSP (treatment check). In contrast, another explanation by Anvari and Lakens was that differences in effectiveness might be caused by a lack of understanding on the meaning and impact of OSP in the research process. This claim, on the other hand, is supported by our results. However, results should be qualified by the fact that we implemented explanations of OSP close to the badges. These text-based specifications also exist in journals using badges (e.g. in the journal 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 text 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). More specifically, the medium-sized negative effect of multiplism on perceived trustworthiness underpins the construct's problematic nature in the context of information processing. Utilizing badges to indicate that science is not just 'opinion' triggered small decreases topic-specific multiplistic beliefs. An important question to clarify would be how sustainable these effects are and if they spill-over onto domain-specific or academic epistemic beliefs when individuals repeatedly notice badges.

In sum, our results further substantiate the assumption that badges produce desirable effects. This is good news for scientists and journal editors since badges are a simple and low-cost method (Kidwell, 2016). Nevertheless, it should also be considered that the meaning and perception of badges is closely tied to the quality standards (as well as their 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 personal website of a researcher might not produce the same effects as badges awarded by journal editors with transparent peer-review standards. Nevertheless, given the promising findings in our study, we conclude that badges hold much potential, which is why we are excited about their further development and implementation.

References

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. (2019). *The Replicability Crisis and Public Trust in Psychological Science* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/vtmpc

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., 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 (Hrsg.), *Personal Epistemology in the Classroom* (S. 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

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

European Commission. (2015). *Study on Open Science. Impact, Implications and Policy Options*. https://ec.europa.eu/research/innovation-union/pdf/expert-groups/rise/study_on_open_science-impact_implications_and_policy_options-salmi_072015.pdf

Fecher, B., & Friesike, S. (2014). Open Science: One Term, Five Schools of Thought. In S. Bartling & S. Friesike (Hrsg.), *Opening Science* (S. 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

Gall, T., Ioannidis, J. P. A., & Maniadis, Z. (2017). The credibility crisis in research: Can economics tools help? *PLOS Biology*, *15*(4), e2001846. https://doi.org/10.1371/journal.pbio.2001846

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for Handling Missing Data. In J. A. Schinka & W. F. Velicer (Hrsg.), *Handbook of Psychology* (2. Aufl., S. 87–114). 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., & Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: The Muenster Epistemic Trustworthiness Inventory (METI). *Plos One*, *10*(10), 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*. 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* (S. 548–573). Guilford Press.

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), 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* (S. 121–144). 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

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

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365. https://doi.org/10.1037/0033-2909.111.2.361

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

Myers, T. A., Kotcher, J., Stenhouse, N., Anderson, A. A., Maibach, E., Beall, L., & Leiserowitz, A. (2017). Predictors of trust in the general science and climate science research of US federal agencies. *Public Understanding of Science*, *26*(7), 843–860. https://doi.org/10.1177/0963662516636040

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 *Information Evaluation* (S. 35–54). ISTE Ltd/John Wiley and Sons Inc.

Pew Research Center (Hrsg.). (2019). *Trust and Mistrust in Americans' Views of Scientific Experts.* https://www.pewresearch.org/science/wp-content/uploads/sites/16/2019/08/PS_08.02.19_trust.in_.scientists_FULLREPORT_8.5.19.pdf

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

Scanlon, M. (2020). *Trust in science and attitudes regarding genetically modified organisms in the U.S. Adults*. Barry University.

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

Sharon, A. J., Yom-Tov, E., & Baram-Tsabari, A. (2020). Vaccine information seeking on social Q&A services. *Vaccine*, *38*(12), 2691–2699. https://doi.org/10.1016/j.vaccine.2020.02.010

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

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* (S. 379–402). MIT Press.

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

van Buuren, S. (2018). *Flexible imputation of missing data* (Second edition). CRC Press, Taylor & Francis Group. https://stefvanbuuren.name/fimd/

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Multivariate Imputation by Chained Equations. *Journal Of Statistical Software*, *45*(3), 1–67. https://doi.org/10.1177/0962280206074463

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 (Proc. InfoVis '11)*, *17*(12), 2223–2230.

https://doi.org/10.1109/TVCG.2011.227

Wingen, T., Berkessel, J., & Englich, B. (2019). *No Replication, no Trust? How Low Replicability Influences Trust in Psychology* [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/4ukq5

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