AERA Open
July-September 2019, Vol. 5, No. 3, pp. 1–18
DOI: 10.1177/2332858419868158
Article reuse guidelines: sagepub.com/journals-permissions
© The Author(s) 2019. http://journals.sagepub.com/home/ero

Smart but Evil? Student-Teachers' Perception of Educational Researchers' Epistemic Trustworthiness

Samuel Merk
University of Tübingen
Tom Rosman
Leibniz Institute for Psychology Information

In-service and preservice teachers are increasingly required to integrate research results into their classroom practice. However, due to their limited methodological background knowledge, they often cannot evaluate scientific evidence firsthand and instead must trust the sources on which they rely. In two experimental studies, we investigated the amount of this so-called epistemic trustworthiness (dimensions expertise, integrity, and benevolence) that student-teachers ascribe to the authors of texts who present classical research findings (e.g., learning with worked-out examples) that allegedly were written by a practitioner, an expert, or a scientist. Results from the first exploratory study suggest that student-teachers view scientists as "smart but evil," since they rate them as having substantially more expertise than practitioners, while also being less benevolent and lacking in integrity. Moreover, results from the exploratory study suggest that evaluativistic epistemic beliefs (beliefs about the nature of knowledge) predict epistemic trustworthiness. A preregistered conceptual replication study (Study 2) provided more evidence for the "smart but evil" stereotype. Further directions of research as well as implications for practice are discussed.

Keywords: *epistemic trustworthiness, epistemic beliefs, student-teachers, teacher education*

Introduction

Teachers all over the world are encouraged to integrate insights from educational research into their everyday practice (Bauer & Prenzel, 2012; Slavin, 2002; Williams & Coles, 2007). However, given their limited knowledge about research methodology, they often cannot evaluate these insights firsthand and must consult secondhand evaluations by asking "Whom do I believe?" instead of "What is true?" (Bromme, Thomm, & Wolf, 2015).

Hence, features such as study design or sample representativeness might be less important when teachers evaluate knowledge claims, whereas criteria such as perceived author expertise or integrity (so-called epistemic trustworthiness; Hendriks, Kienhues, & Bromme, 2015) become pivotal. Research into the predictors of epistemic trustworthiness, however, is still in its infancy (for the first experimental attempts, see, e.g., Hendriks, Kienhues, & Bromme, 2016a; Thon & Jucks, 2017). Do teachers and student-teachers see certain sources as more credible than others? Does the trustworthiness they ascribe to certain sources depend on their beliefs regarding the nature of scientific knowledge (Hofer & Bendixen, 2012; so-called epistemic beliefs; Hofer & Pintrich, 1997)? In view of this research gap and considering that such predictors may provide valuable insights into how we should approach teacher education, we investigated student-teachers' perceived epistemic trustworthiness of educational researchers and how it relates to their epistemic beliefs in an exploratory pilot study and a preregistered main study. Before we describe the two studies in detail, we will provide some background information about the constructs they entail.

Epistemic Trustworthiness and Epistemic Beliefs

Epistemic Trustworthiness

In-service teachers and teacher education students are confronted with vast amounts of information about teaching that stem from a multitude of information sources. For example, they may read an expert blog that introduces a new digital classroom tool, consult their colleagues' opinions on a certain teaching method, or skim a newspaper article on school reforms. Moreover, in line with current calls for more evidence-based practice in education (e.g., Munthe & Rogne, 2015), in-service and student-teachers increasingly are required to inform themselves using science-based information sources (e.g., empirical studies or scientific textbooks).

Before using any such information for their everyday practice, it is vital that teachers evaluate its veracity through its logical coherence and cohesiveness (Bromme, Kienhues, Porsch, Bendixen, & Feucht, 2010). Regarding scientific knowledge, however, two aspects make this endeavor particularly challenging (Hendriks, Kienhues, & Bromme, 2016b). First, since it relies on axioms and only offers "degrees of confirmation" (Popper, 1954), scientific knowledge always encompasses some degree of (epistemic) uncertainty (Retzbach, Otto, & Maier, 2016; Sinatra, Kienhues, & Hofer, 2014). Hence, finding out what is "true" is not as easy as it might seem. Second, modern science is highly specialized and has developed a "social infrastructure of knowledge in which there are divisions of cognitive labor and sophisticated mechanisms for recognizing appropriate experts and knowing when and how to defer to them" (Keil, 2010, p. 828). Due to this division of cognitive labor and since teachers (and student-teachers) usually do not have much research expertise, firsthand (i.e., direct) evaluations of the veracity of scientific knowledge are often unfeasible. Therefore, secondhand evaluations (Bromme et al., 2010) come into play, as individuals no longer directly assess the veracity of knowledge claims, but they assess the credibility and trustworthiness of the sources from which this knowledge originates.

When individuals assess the trustworthiness of different sources, they refer to specific source information features (Stadtler, Scharrer, Macedo-Rouet, Rouet, & Bromme, 2016). In this regard, the professional background of the source is of particular importance, as it helps individuals "decide whether the source possesses expertise that is pertinent to his or her current problem" (Stadtler et al., 2016, p. 709). A caveat when individuals use such source information, however, is that biased prior beliefs on specific sources may affect their information behavior considerably, to a point at which they refrain from using specific source types (e.g., science-based sources) altogether. Regarding teacher education, this is an important question for everyone aiming to promote evidence-based practice: If teachers mistrust science-based information, they likely will not use such knowledge in their teaching and instead rely on experiential and anecdotal evidence. This may, in turn, affect their teaching considerably, especially considering that in the education domain, a multitude of other readily available sources are available—for example, colleagues' knowledge and expertise or personal experiences (Buehl & Fives, 2009). Therefore, we see it as crucial to investigate the "epistemic trust" that (student) teachers attribute to scientific sources (e.g., findings from educational studies)—especially in contrast with their trust in nonscientific findings (e.g., teachers).

Several studies from the realm of science communication (Cummings, 2014; Hendriks et al., 2015; Peters, Covello, & McCallum, 1997) and other fields (e.g., Landrum, Mills, & Johnston, 2013; Mayer, Davis, & Schoorman, 1995) suggest operationalizing epistemic trust in three dimensions: expertise, benevolence, and integrity, which can be applied to scientific as well as nonscientific sources (e.g., trust in the expertise of teachers vs. educational reseachers). A source exhibits high expertise if it is highly informed, intelligent, and qualified. Benevolent sources are interested in the greater good of others, and sources with integrity respect norms and values in great measure. This conceptualization is particularly fruitful, as it allows for a more fine-grained investigation into the different aspects of epistemic trust in educational science than in studies that analyze the trustworthiness of scientific practices in general (Collins, 2009; Nadelson & Hardy, 2015; Wynne, 2006).

In sum, these deliberations lead us to our first (exploratory) research question:

> *Research Question 1:* What amount of epistemic trust (expertise, benevolence, and integrity) do student-teachers attribute to different sources of educational knowledge?

Extant research from related fields has shown that teacher education students have a rather negative attitude toward scientific knowledge in general, at least when considering educational disciplines; for example, they view general pedagogical knowledge as too abstract and theoretical (Gitlin, Barlow, Burbank, Kauchak, & Stevens, 1999; Sjølie, 2014; van der Linden, Bakx, Ros, Beijaard, & Vermeulen, 2012). While we tended to assume that student-teachers also *mistrust* scientific knowledge, in line with such research, we did not formulate specific confirmatory hypotheses prior to investigating this research question by means of existing data. Therefore, the present article is divided into a pilot study (Study 1) with an exploratory nature and a main study (Study 2), which was preregistered to ensure confirmatory research (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

Epistemic Beliefs

Epistemic trustworthiness focuses on the expertise, benevolence, and integrity that individuals ascribe to specific sources of knowledge (e.g., scientists or practitioners). Epistemic beliefs, in contrast, encompass explicit and implicit beliefs about knowing and knowledge, and can be defined as "an identifiable set of dimensions of beliefs, organized as theories, progressing in reasonably predictable directions, activated in context, operating as epistemic cognition" (Hofer, 2001, p. 377). This definition already highlights that researchers study epistemic beliefs under various notions and perspectives. In the earlier years of epistemic belief research, researchers primarily adopted a developmental perspective that was pioneered by Perry (1970), who interviewed college students and deduced a scheme to describe the development of their epistemic beliefs. According to Perry (1970), epistemic beliefs develop in nine steps over four developmental stages, beginning with a dualist stage, in which individuals view knowledge as absolute and certain. In the next stage, called multiplism, individuals perceive knowledge as created by the human mind, whereby assertions become more like opinions than facts. Individuals with multiplistic beliefs tend to very similar validity claims regarding different types of arguments (e.g., scientific vs. episodic evidence) by emphasizing that "knowledge is subjective, uncertain, and justified by personal preferences and judgments" (Barzilai & Eshet-Alkalai, 2015). In the two highest stages, relativism and commitment within relativism, individuals accept the tentativeness of scientific knowledge and its origin in the human mind but believe that knowledge is susceptible to evaluation. Individuals with evaluativistic beliefs hence focus on evaluating the validity claims of knowledge assertions instead of neglecting the possibility of this evaluation (multiplistic beliefs) or assuming that knowledge assertions are either true or false (absolutistic beliefs). Perry's (1970) model was adopted and modified by many researchers (e.g., Greene, Azevedo, & Torney-Purta, 2008; Krettenauer, 2005; Kuhn, Cheney, & Weinstock, 2000), whose models vary in terms of the conceptualization, number, and labeling of respective stages. However, even though the developmental perspective is still used today (Muis, Bendixen, & Haerle, 2006), the assumption that several aspects of epistemic beliefs develop simultaneously when individuals move from one stage to another has been questioned several times (e.g., Hofer & Pintrich, 1997). Over the years, this has led to the so-called dimensional perspective, under which different subdimensions of epistemic beliefs can be distinguished. The most prominent dimensional framework was suggested by Hofer and Pintrich (1997) and includes two dimensions of beliefs about knowledge (simplicity and certainty) and two dimensions of beliefs about knowing (source and justification). However, dimensional frameworks also have been criticized, mainly for their conceptual muddiness in defining the dimensions' extreme poles. This has led to the development of several integrated frameworks (Barzilai & Eshet-Alkalai, 2015; Greene et al., 2008; Peter, Rosman, Mayer, Leichner, & Krampen, 2016; Rule & Bendixen, 2010), which posit that epistemic beliefs develop within multiple dimensions over diverse stages (e.g., absolutism, multiplism, and evaluativism; Muis et al., 2006).

As for relationships between epistemic beliefs and epistemic trust, reflections of epistemological entities form a *condicio sine qua non*, at least for some parts of epistemic trust: If one is epistemologically pessimistic (i.e., believing that reality is not accessible to scientists), it does not make sense to attribute high expertise to these scientists. Or vice versa: Assuming that a researcher is competent, benevolent, and has integrity is, by definition, consistent with evaluativistic epistemic beliefs. If scientific knowledge *does not* consist of arbitrary opinions, but instead is susceptible to evaluations of its assertions through the scientific community, then at least most scientists are likely trustworthy. Hence, we formulated the following second research question:

> *Research Question 2:* Can domain-specific beliefs about educational research predict epistemic trust in sources of assertions from educational research?

For the same reason as outlined above, this research question again was investigated in one exploratory study (Study 1) and one confirmatory, hypothesis-testing study (Study 2).

Present Studies

In this section, we present two studies that investigate the research questions mentioned above and that are repeated here for readers' convenience: (1) *What amount of epistemic trust (expertise, benevolence, and integrity) do student-teachers attribute to different sources of educational knowledge?* (2) *Can domain-specific beliefs about educational research predict epistemic trust in sources of assertions from educational research?* Study 2 was designed to conceptually replicate the results from Study 1, and its hypotheses were preregistered to ensure confirmatory, hypothesis-testing research (Nosek et al., 2015).

Study 1: Exploratory Study

Procedure and Materials

A challenge when investigating students' trust in different sources is that trust ratings are confounded by the types of information that are usually associated with specific sources. For example, when asking study participants about their trust in scientific sources, they might, in reality, state their opinions about educational theories—or their responses might at least be biased by such opinions. When asking about their trust in practitioners, they might think about one specific colleague who cherishes controversial teaching methods. This is even more problematic considering that a large proportion of variance in students' epistemic beliefs is located at the topic level, meaning their beliefs strongly vary with regard to different topics and contexts (Merk, Kelava, Schneider, Syring, & Bohl, 2017; Merk, Rosman, Muis, Kelava, & Bohl, 2018; Trautwein & Lüdtke, 2007).

To circumvent this issue and explore whether student-teachers' trust in scientific knowledge indeed depends on the *source* of such knowledge, we developed text materials that were invariant in content (i.e., contained the same body of knowledge), but varied in sources. To achieve this, five researchers independently collected curricular valid educational research topics (e.g., specific theories, effects, or findings), then discussed and evaluated the representativeness of these topics for the domain of educational research and their

TABLE 1

Epitomes From the Intervention Texts

Topic	Practitioner statement	Scientific study
Bullying/mobbing	*During my internship in a middle school, I was shocked about how much bullying has spread since my school days.* When I write about bullying, I mean **intentional and repeated negative behavior of one or more students against another student**. . . . *My own experience and the experience of colleagues* show that **about every fourth middle school student and every tenth high school student is being bullied at school**. . . .	*Our working definition of bullying pertains to Olweus (2010)*, who described it as **intentional and repeated negative behavior of one or more students against another student**. . . . *Fellow researchers have already found out* that **about every fourth middle school student and every tenth high school student is being bullied at school** *(Whitney & Smith, 1993)*. . . .

appropriateness for experimental manipulation. Four topics were chosen (“learning from worked-out examples,” “cognitive theory of multimedia learning,” “bullying/mobbing,” and “classroom size effects on achievement”). Subsequently, invariant text components were created that contained the core information in terms of descriptions of the theories, effects, or findings in question. Finally, the context information pertaining to the (alleged) source of the knowledge was added by means of additional sentences. To enhance the study’s internal validity, three of the five researchers were randomly assigned to the two writing steps and had to fulfill criteria concerning text length (130 words < text length < 200 words) and text complexity (50 < Läsbarhetsindex [Readability Index] = LIX < 65). The full body of material can be found in online Supplemental Appendix 1. Table 1 provides epitomes of the texts.

Design

Study 1 used a between-person design. After responding to some demographic questions and filling out an epistemic beliefs inventory (see “Measurements” section below), every participant read four texts (with four different topics; see “Procedure and Materials” section above), which all contained, for each participant, the same alleged source (“practitioner,” “expert,” or “scientific study”). On reading each text, the participants additionally responded to some text-specific questions for purposes of another study (Merk, Rosman, Ruess, Syring, & Schneider, 2017). After having read all four texts, participants responded to an item battery containing among others the treatment check and the epistemic trustworthiness inventory.

Sample

Participants (N = 365, 243 females, 51% in the first two semesters) were recruited through slides during lectures and informed that participation was voluntary and could be stopped at any time, that each participant was allowed to participate in a lottery of five vouchers worth €50, and that the study would take about 40 minutes. Data collection was conducted in paper–pencil format. The questionnaires were transcribed to raw data through automated scanning software.

Measurements

Treatment Check. To ensure that the readers perceived the texts’ sources as intended, we asked them to rate authors’ characteristic occupational activities of their respective texts (question prompt: *“What do you think: How frequently do the authors of your texts engage in the following activities?”* sample item practitioner: “*teaching at school*”; sample item expert statement: “*give advice to schools*”; sample item scientific study: “*investigating data*”; response format: 6-point Likert-type scale; all items can be found in online Supplemental Appendix 2). A multiple indicators, multiple causes (MIMIC; Jöreskog & Goldberger, 1975) model with the source-specific activities as indicators and two dummy variables encoding the three sources as causation (see Figure 1) was fitted to the data. According to widespread benchmarks (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004), the model fit was not perfect, but suitable for the purposes of a treatment check (χ^2 = 121.8, df = 35, CFI [comparative fit index] = .954, TLI [Tucker–Lewis index] = .929, RMSEA [root mean square error of approximation] = .84, SRMR [standardized root mean square residual] = .058). Parameter estimates indicated that the participants strongly differentiate between “practitioner” sources and the other two varieties, but rather weakly distinguish between “expert” and “scientific study” sources. Since we specified τ-congeneric measurement models, we assessed the reliability (internal consistency) of the treatment check scales with McDonald’s ω (Dunn, Baguley, & Brunsden, 2013). Reliability was good for the practitioner (ω = .846, 95% confidence interval [CI] [.815, .878]) and scientist scales (ω = .872, 95% CI [.845, .900]), but questionable for the expert scale (ω = .578, 95% CI [.494, .662]).

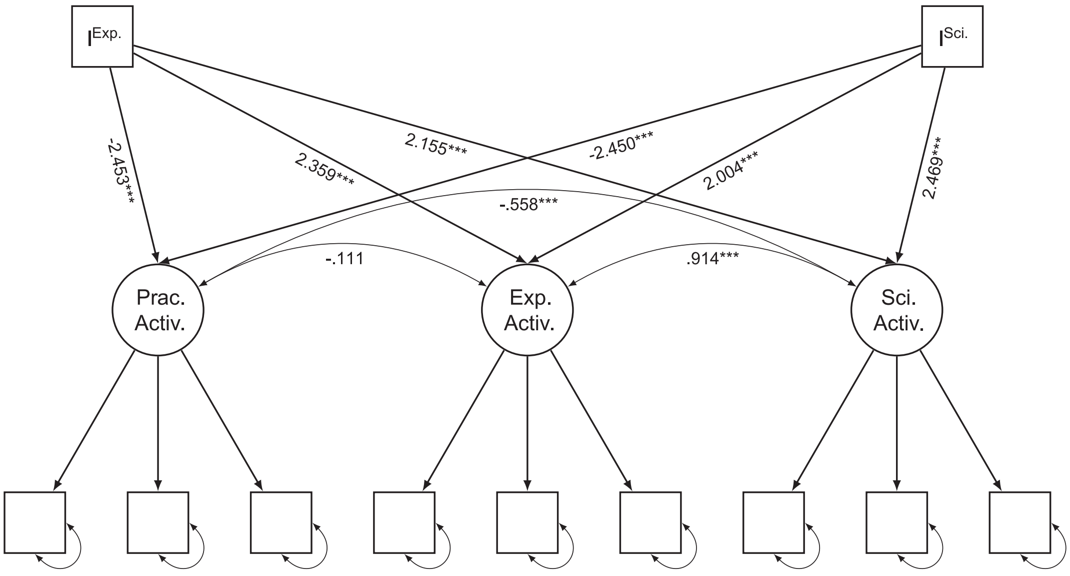

FIGURE 1. *MIMIC model of the treatment check.*
Note. MIMIC = multiple indicators, multiple causes; $I^{Exp.}$ = dummy variable for source "expert"; $I^{Sci.}$ = dummy variable for source "scientific study"; reference category = source "practitioner"; Sci. = author of scientific study; Exp. = expert; Pract. = practitioner; Activ. = activity. Regression paths from dummy variables are *y*-standardized.
***$p < .001$.

Epistemic Trustworthiness. The Muenster Epistemic Trustworthiness Inventory (METI; Hendriks et al., 2015) was used to assess the extent of epistemic trustworthiness that student-teachers attribute to the different sources. METI is constructed as a semantic differential and consists of the three dimensions: "expertise," "benevolence," and "integrity" (see "Measurements" section). To investigate the scales' construct validity, we first performed a confirmatory factor analysis (CFA). We specified a model with three factors, τ-congeneric measurement models, and three freely estimated residual covariances (based on modification indices), which resulted in good model fit (χ^2 = 194.0, *df* = 71, CFI = .956, TLI = .943, RMSEA = .072, SRMR = .061). Reliability (assessed by McDonald's ω) was good for all three scales (expertise: ω = .884, 95% CI [.855, .912]; benevolence: ω = .868, 95% CI [.843, .893]; integrity: ω = .838, 95% CI [.784, .892]).

Epistemic Beliefs. We used a domain-specific adaptation of the German FREE questionnaire (FREE; Krettenauer, 2005; Merk, Rosman, et al., 2017), an instrument based on Kuhn and Weinstock's (2002) framework, to assess the level of development of domain-specific epistemic beliefs. Using a scenario-based approach (e.g., Händel, Artelt, & Weinert, 2013), the instrument describes 13 well-known educational research issues (e.g., "It is repeatedly discussed whether grade retention is actually useful or should be abolished") and prompts participants to indicate their (dis)agreement with three statements representing the three stages of epistemic development for each presented issue (6-point Likert-type scale; sample statement for the absolutism stage: *Either grade retention is useful or not! Educational researchers should unequivocally clarify this in the future*; multiplism: *The expressions for "grade retention" are mere conjecture; no one can really know which factors contribute to school achievement*; evaluativism: *Even though the experts disagree, both may present more or less good reasons for their conceptions*). Krettenauer (2005) suggests computing a so-called d-index (d = eval − 0.5 * mult − 0.5 * abs) for each issue/scenario. We followed this suggestion and computed a CFA on the scales' 13 d-indices with a τ-congeneric measurement model and two freely estimated residual covariances (selected by modification indices), which showed good data adaptation to the model (χ^2 = 98.8, *df* = 63, CFI = .930, TLI = .913, RMSEA = .043, SRMR = .047). Internal consistency of the d-index was satisfactory (McDonald's ω = .75, 95% CI [.71, .80]).

Statistical Analyses

We decided to use multiple regression analysis (Fox & Weisberg, 2011) to answer both research questions. Multiple regression is a so-called complete data method when estimated with least squares. Since simple approaches such as listwise deletion potentially result in biased parameter estimates or lower statistical power (Rubin, 1976; Schafer &

TABLE 2
Means and Standard Deviations of the METI Dimensions in Study 1

METI dimension	Source	*M*	*SD*
Benevolence	Expert	5.23	0.89
	Practitioner	5.76	0.79
	Scientific study	5.23	0.84
Expertise	Expert	5.32	0.92
	Practitioner	5.26	0.96
	Scientific study	5.65	0.72
Integrity	Expert	5.40	0.79
	Practitioner	5.79	0.70
	Scientific study	5.42	0.74

Note. METI = Muenster Epistemic Trustworthiness Inventory.

Graham, 2002), we had to explicitly deal with missing data. Hence, we used multiple imputation on our raw data (0% to 16.6% missing data) by means of chained equations (Azur, Stuart, Frangakis, & Leaf, 2011) using functions provided by the R-package "mice" (van Buuren & Groothuis-Oudshoorn, 2011). We subsequently estimated the regression models separately on all resulting (30) complete data sets and combined the results using the formulae provided by Rubin (1976).

Results

Initially, we inspected the data descriptively (see Table 2) and graphically (see Figure 2). To answer Research Question 1, we recoded the independent variable "source" into two dummy variables, I^{Expert} and $I^{\text{Scientific Study}}$ (reference category: practitioner), and conducted a multiple regression analysis with the *z*-standardized dependent variables "expertise," "benevolence," and "integrity." Hence, the slope parameters of these dummy variables can be interpreted as estimates of differences between the group specified in the respective dummy variable and the reference group. As Table 3 shows, there were moderate to large differences between the practitioner group and the scientific study group in all three dimensions of epistemic trustworthiness: The authors of scientific studies are perceived not only as being less benevolent, with less integrity, but also as having more expertise (see Table 3). All these effects were statistically significant. This was, likewise, the case for differences between the practitioner and expert sources, but only for the dimensions "benevolence" and "integrity" and not for the dimension "expertise" (see Table 3).

To investigate Research Question 2, we added epistemic beliefs (d-index) to the former models. As can be seen in Table 3, parameter estimates indicated small effects on all dimensions of epistemic trustworthiness. We interpret this as preliminary evidence for an association of epistemic development and epistemic trustworthiness: Student-teachers who believe that (scientific) educational knowledge consists not so much of "absolute facts" (absolutism) and "arbitrary opinions" (multiplism), but more of assertions whose validity can be evaluated (evaluativism), tend to show higher epistemic trustworthiness on all three dimensions ("expertise," "benevolence," and "integrity").

Interim Discussion of Study 1

Study 1 investigated (1) whether student-teachers tend to trust sources of assertions from educational research (practitioner, expert, and scientific study) differentially and (2) whether their amount of epistemic trust in these sources can be predicted by their epistemic beliefs. Regarding the first question, we found what one may call a "smart but evil" stereotype, as the authors of scientific studies (i.e., scientists) are perceived not only as less benevolent, with less integrity, but also as having more expertise in contrast to practitioners. This is an intriguing finding, as it suggests that student-teachers hold a kind of distrust in scientists ("Scientists have the expertise to find answers, but they do not really want to!"). Regarding Research Question 2, we found small effects from an evaluativistic view of (scientific) educational knowledge on the epistemic trustworthiness of this knowledge's source.

However, despite several methodological strengths (e.g., the experimental variation of sources or the high construct validity of the measurements), there are three particular limitations that motivated us to undertake a conceptual replication of these findings (Simons, 2014) in the form of a preregistered (Nosek et al., 2015; van 't Veer & Giner-Sorolla, 2016) and, therefore, clearly confirmatory (Wagenmakers et al., 2012) study. First, in the field of epistemic beliefs, there is an emerging call for disentangling epistemic beliefs (and related constructs) of varying specificity and different contexts (Buehl & Alexander, 2006; Merk et al., 2018; Muis et al., 2006). However, Study 1 neglects this differentiation. In fact, our participants read *topic-specific* assertions stemming from *different sources*, rated the epistemic trustworthiness *aggregated for all four texts*, and responded to a *domain-specific* measurement of epistemic beliefs.

Second (and this seems somewhat close but is in fact substantially different from the first point), we want to highlight that Study 1 only investigated source-specific differences in epistemic trustworthiness and its relation to epistemic beliefs in a *between-person design*. While we judge this as appropriate for a first exploratory study, a large amount of research empirically and conceptually has shown that there is substantial variation in epistemic beliefs within persons (i.e., one individual may have very different beliefs regarding different topics or domains) *and* between persons (i.e., individuals stemming from different domains may have different beliefs regarding the same topic or domain; Buehl & Alexander,

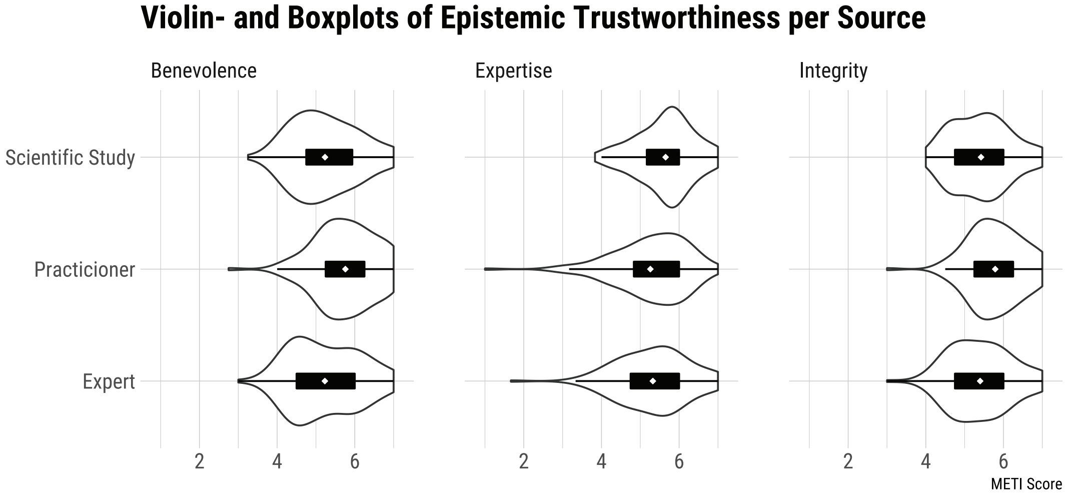

FIGURE 2. *Violin- plots and boxplots of the results (Study 1).*
Note. Rhombs are depicting arithmetic means.

TABLE 3
Pooled Results From Multilevel Regression

	Dependent variable					
	Expertise		Benevolence		Integrity	
	M1	M1b	M2	M2b	M3	M3b
Intercept	−.18 (.09)	−.18 (.09)	.40*** (.09)	.40*** (.09)	.33*** (.09)	.33*** (.09)
I^{Expert}	.08 (.13)	.08 (.13)	−.59*** (.13)	−.59*** (.12)	−.50*** (.13)	−.51*** (.13)
$I^{Sci.\ Study}$	.45** (.13)	.44** (.13)	−.59*** (.12)	−.59*** (.12)	−.47*** (.13)	−.47*** (.13)
d-Index		.12* (.05)		.11* (.05)		.15** (.05)
R^2	.039	.052	.076	.088	.052	.075

Note. I^{Expert} = dummy coded indicator variable for source “expert”; $I^{Sci.\ Study}$ = dummy coded indicator variable for source “scientific study.”
**p* < .05. ***p* < .01. ****p* < .001.

2001, 2006; Hofer, 2006; Limón, 2006; Merk, Kelava, et al., 2017; Muis, 2004; Trautwein, Lüdtke, & Beyer, 2004; Trautwein & Lüdtke, 2007). Thus, to ensure more detailed conclusions, we see it as crucial to assess source-specific differences in epistemic trustworthiness within and between persons *simultaneously* as (despite between-person differences) there might be substantial within-person variations of the “smart but evil” stereotype. For example, it is very conceivable that student-teachers view scientists as having much more expertise regarding the “cognitive theory of multimedia learning” topic, but view practitioners as having nearly equal expertise in the topic of “bullying/mobbing” while viewing scientists as having moderately more expertise overall.

Third, as mentioned above, we did specify the research questions before analyzing the data from Study 1, but we did not have specific hypotheses and no detailed *a priori* analysis plan. Hence, due to its exploratory nature, the evidence gathered in Study 1 is less robust than it might seem (Chambers, 2017). Therefore, we used the theoretical background and empirical results from Study 1, considered its methodological strengths and weaknesses, and derived a set of specific hypotheses that were preregistered (Merk & Rosman, 2019) and tested along a (likewise preregistered) detailed data analysis plan in Study 2. Study 2 used the same materials as Study 1 and investigated the same research questions but drew on an enhanced design. Therefore, it should be viewed as an attempt of a conceptual replication of Study 1 (Simons, 2014).

Study 2: Confirmatory Study

Research Questions and Hypotheses

Research Question 1. The first research question focuses (for both studies) on the amount of epistemic trust that student-teachers attribute to different sources of educational knowledge. In Study 1, we found what one may call a “smart but evil” stereotype: The authors of scientific studies (i.e., scientists) were perceived as less benevolent, with less integrity, but having more expertise in contrast to practitioners.

To replicate these findings conceptually, we suggest the following confirmatory hypothesis for Study 2:

> *Hypothesis 1:* Student-teachers ascribe less integrity and benevolence, but more expertise, to scientific information sources in contrast to practitioner sources.

Research Question 2. Our second research question aims (for both studies) at investigating whether epistemic beliefs about educational research can predict epistemic trust in sources of assertions from educational research. As already outlined above, we see evaluativistic, domain-specific, epistemic beliefs as a necessary condition for epistemic trust, thereby suggesting the following hypothesis:

> *Hypothesis 2a:* Evaluativistic, domain-specific, epistemic beliefs are positively related to ascriptions of integrity, benevolence, and expertise.

To overcome the conceptual weakness of Study 1 concerning the blurred levels of specificity (see above), we added an analogous, but more specific, hypothesis:

> *Hypothesis 2b:* Multiplistic topic-specific, epistemic beliefs are negatively related to ascriptions of integrity, benevolence, and expertise.

Design

To enhance Study 1 while simultaneously replicating it conceptually, we planned to test both hypotheses as within- and between-persons effects simultaneously (see the "Discussion" section for Study 1). Therefore, it must be ensured that (1) every participant reads assertions stemming from different sources and that the number of sources per participant is balanced out for each participant (within-person component), (2) no participant reads the same assertion more than once, and (3) combinations of topics and sources, as well as the sequence of topics and sources, cannot confound our results. Since the main focus of the present study is the distinction between scientific and practitioner sources, and since the manipulation check regarding the "expert source" level indicated some problems, we decided to reject this "intermediate" level. This also reduces cognitive load on the participants, allowing us to include all four topics from Study 1 seamlessly without running into randomization problems.

To randomize the different sources and topics, we first created a Latin square of the four topics to achieve (incomplete) counterbalance (DePuy & Berger, 2014) in the topic factor. Subsequently, we repeated this procedure six times and addressed all possible sequences of the two texts regarding the sources "practitioner" and "scientific study" (see Table 4) to counterbalance this factor as well.

Procedure and Materials

All materials in Study 2 were identical to those used in Study 1, but, corresponding to the different design of Study 2, the experimental procedure differed: Participants were assigned randomly to 1 of 24 different questionnaires (see Table 4 and online Supplemental Appendix 4 for the questionnaries of Study 2). To achieve this, we used an urn model (e.g., Wei, 1978) and true random numbers obtained at www.random.org to ensure that every questionnaire is filled out with the same frequency. Just like in Study 1, the participants first filled out the FREE questionnaire (domain-specific measurement of epistemic beliefs), then went through the four topics (sequence and sources of the assertions, depending on the questionnaire). After each topic, they filled out the METI, a topic-specific multiplism scale (see "Instruments" section below), and the items of the treatment check. Finally, the participants were asked for some demographic data.

Measurements

As mentioned above, we used the FREE and METI to assess domain-specific epistemic beliefs and epistemic trustworthiness, respectively. Both instruments are described in the "Methods" section for Study 1 and are provided at full length in online Supplemental Appendix 2, where all other instruments can be found as well. Additionally, we measured topic-specific multiplism by means of the "topic-specific multiplism" scale (4-point Likert-type scale), which was developed by decontextualizing the FREE's multiplism items (Merk, Schneider, Syring, & Bohl, 2016) and has demonstrated good psychometric properties in several studies (Merk, Kelava, et al., 2017; Merk, Rosman, et al., 2017)

Statistical Analyses

Psychometric Properties. The psychometric properties of the only between-person measurement (FREE) were evaluated just like in Study 1. We first ran a CFA with τ-congeneric measurement models and allowed for residual covariances identified by modification indices (Standardized Expected Parameter Change; Whittaker, 2012). Reliability (internal consistency) was assessed using McDonald's ω. Just like in Study 1, indicators of acceptable/good fit were CFI and TLI values that exceed .90/.95, RMSEA values lower than .10/.06, and SRMR values inferior to .08/.05 (Browne & Cudeck, 1992; Hu & Bentler, 1999).

The factorial validity of the within-person measurements METI, topic-specific multiplism, and treatment check was assessed using multilevel confirmatory factor analysis (MCFA; Mehta & Neale, 2005; B. O. Muthén, 1994). To do

TABLE 4

Design of Study 2: Counterbalanced Sequences of Topics and Sources

		Text position			
		1	2	3	4
1st Latin square	Questionnaire 1	bm prac	cs sci	we sci	cm prac
	Questionnaire 2	we prac	cm sci	bm sci	cs prac
	Questionnaire 3	cs prac	bm sci	cm sci	we prac
	Questionnaire 4	cm prac	we sci	cs sci	bm prac
2nd Latin square	Questionnaire 5	we prac	cm sci	bm prac	cs sci
	Questionnaire 6	bm prac	cs sci	we prac	cm sci
	Questionnaire 7	cm prac	bm sci	cs prac	we sci
	Questionnaire 8	cs prac	we sci	cm prac	bm sci
3rd Latin square	Questionnaire 9	bm prac	cm prac	we sci	cs sci
	Questionnaire 10	cs prac	we prac	cm sci	bm sci
	Questionnaire 11	cm prac	bm prac	cs sci	we sci
	Questionnaire 12	we prac	cs prac	bm sci	cm sci
4th Latin square	Questionnaire 13	bm sci	cs prac	cm sci	we prac
	Questionnaire 14	cm sci	we prac	bm sci	cs prac
	Questionnaire 15	we sci	cm prac	cs sci	bm prac
	Questionnaire 16	cs sci	bm prac	we sci	cm prac
5th Latin square	Questionnaire 17	cs sci	cm prac	bm prac	we sci
	Questionnaire 18	we sci	bm prac	cs prac	cm sci
	Questionnaire 19	cm sci	cs prac	we prac	bm sci
	Questionnaire 20	bm sci	we prac	cm prac	cs sci
6th Latin square	Questionnaire 21	bm sci	cm sci	cs prac	we prac
	Questionnaire 22	cm sci	bm sci	we prac	cs prac
	Questionnaire 23	cs sci	we sci	cm prac	bm prac
	Questionnaire 24	we sci	cs sci	bm prac	cm prac

Note. we = learning from worked-out examples; cm = cognitive theory of multimedia learning; bm = bullying/mobbing; cs = classroom size effects on achievement; prac = practicioner; sci = scientific study.

so, we specified MCFA models with τ-congeneric measurement models at each level, whereby we addressed the same cutoff values for model-fit evaluation for MCFA as in the single-level case (FREE; see above), but calculated SRMR separately for each level, whereby we defined the $\text{SRMR}_{\text{Between}}$ values smaller than .10/.05, indicating acceptable/good fit.

Confirmatory Analysis Plan. For all statistical tests, the cutoff for statistical significance was a *p* value of .05. Our design produces clustered data, as each individual is subjected to the within-person measurements four times (once for each text). Hence, multilevel regression is an appropriate method for modeling within-person variations and between-person differences simultaneously (Gelman & Hill, 2007; Raudenbush & Bryk, 2002). Research Question 1 deals with source effects on epistemic trustworthiness, which we examined on both within-person and between-person levels.

On the within-person level, we were interested in whether METI scores vary intra-individually depending on source differences (practitioner vs. scientific study) between the four texts that each participant read. Therefore, we specified random-intercept models with a dummy-coded indicator variable (practitioner: value 1; scientific study: value 0) indicating the source as a predictor of each of the three dimensions of the METI (in three models named M1a, M1b, and M1c). These effects were tested with *t* tests on the fixed effects and likelihood ratio tests (LRTs, as opposed to an intercept-only model; Hox, 2010).

Regarding Research Question 1, we were interested in whether differences exist *between individuals* in METI scores on the same topics, depending on our source manipulation of the respective texts. Since each participant responded to exactly two “practitioner” and two “scientific study” texts (see Table 4), this must be tested separately for each topic. Hence, the source (here coded as a dummy variable) was regressed on the dimensions of epistemic trustworthiness in four single-level path models —one for each topic (Models M2a–M2d). These between-person effects were statistically evaluated using *t* tests for the (standardized) path coefficients. For M2a to M2d, missing values were handled using a model-immanent

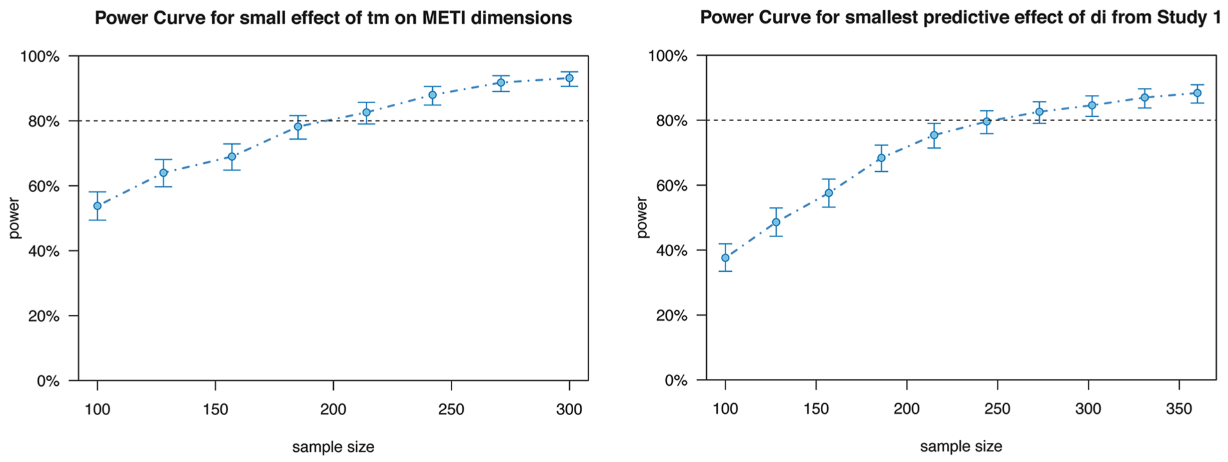

FIGURE 3. *Power obtained by Monte Carlo studies for predictive effects of topic-specific multiplism and the d-index.*

approach using full information maximum likelihood estimation (Finkbeiner, 1979)

To answer Research Question 2 regarding both within- and between-person effects, we extended models M1a, M1b, and M1c with topic-specific multiplism as a within-person level predictor and the FREE's d-index as a between-person level predictor. The resulting models were labeled M3a, M3b, and M3c, respectively. Topic-specific multiplism was centered on the cluster means, so that the predictive effects of the d-index can be interpreted as effects on the person-specific means of the respective epistemic trustworthiness dimension (Trautwein & Lüdtke, 2009): Hence, one can view the effects of the source and topic-specific multiplism as within-person effects simultaneously modeled with the between-person effect of the d-index. Fixed effects were tested using *t* tests, along with LRTs of corresponding nested models (e.g., M1a vs. M3a).

Handling of Missing Data. As we used paper-and-pencil questionnaires, it was very likely that we would have to deal with missing data. To avoid problems associated with "naïve" handling of missing data (e.g., listwise deletion), we imputed the data by means of multiple imputation under a joint modeling perspective (Schafer & Yucel, 2002) using the R-package "pan" (Zhao & Schafer, 2016), as this package is specialized for the multiple imputation of multilevel data. Just like in Study 1, we tested the models described above on each complete data set and combined the results using the rules proposed by Rubin (1987).

Sampling Plan

Recruitment. Study participants were recruited from several teacher education courses at the University of Tübingen, Germany. Inclusion criteria were that chosen participants (1) were teacher education students at the University of Tübingen and (2) had not participated in Study 1. Adherence to these inclusion criteria was ensured by respective promotion and assessment of the coherent covariates. Participation was voluntary and took place during class time. As an incentive, all participants could participate in a lottery for vouchers worth €50.

Power Analysis. Evaluating the statistical power of the multilevel regression and single-level path models was a challenge because it depends on several factors that can be determined only empirically (e.g., variable distribution or amount of missing data). To anticipate the statistical power of the models in Study 2, we used the results from Study 1, using the Monte Carlo approach (L. K. Muthén & Muthén, 2002), in which a large set of sample data is drawn from a hypothesized population model, and parameters and standard errors are estimated for each of the sample data sets, which are then averaged.

To evaluate the power of the planned multilevel regression analyses, we set up an artificial data set corresponding to our design and subsequently sampled values for the d-index (which are independent of the experimental condition; see "Design" section above) from Study 1. In the next step, we simulated a sample of topic-specific multiplism, considering the effects of each condition's source and topic, as well as the association to the d-index using an R package named "simr" (Green & MacLeod, 2016). Finally, we carried out the simulation study with a conservative setting: For the predictive effect of theory-specific multiplism, we assumed a "small" effect following Cohen's benchmarks (Cohen, 1988). For the FREE's d-index, we used the smallest effect size from Study 1. As can be seen in Figure 3, the traditional benchmark of power >.80 is achieved for both effects at a sample size of approximately 264 (11 individuals per questionnaire).

To anticipate the statistical power of the planned models M2a to M2d, we again used a Monte Carlo approach based on the data from Study 1. As only two sources will be used

TABLE 5

Psychometric Properties of the Measurements Used in Study 2

Measurement	$\chi^2(df)$, TLI/CFI, $\mathrm{SRMR}_{(\mathrm{Between/Within})}$	Dimension	McDonald's ω (minimum, maximum)
FREE	97.47 (63), .902/.921, .049	d-Index	.730
METI	861.46 (147), .930/.943, .076/.095	Expertise	(.914, .935)
		Benevolence	(.892, .908)
		Integrity	(.880, .909)
Treatment check	53.466 (18), .991/.995, .012/.093	Practitioner activities	(.907, .945)
		Scientist activities	(.946, .962)
Topic-specific multiplism	32.30 (4), .902/.967, .024/.087	Multiplism	(.787, .842)

Note. TLI = Tucker–Lewis index; CFI = comparative fit index; SRMR = standardized root mean square residual; FREE = German FREE questionnaire; METI = Muenster Epistemic Trustworthiness Inventory. If measurements were multiply applied within persons, McDonald's ω was computed separately for each topic. The corresponding minimum and maximum values are given in the table.

in Study 2 (practitioner and scientific study), we subsetted the data from Study 1 accordingly, ran a path model that predicts METI dimensions by a dummy variable of the source (1 = scientific study, 0 = practitioner), and used the resulting parameters as population parameters for the Monte Carlo study (see online Supplemental Appendix 3). The results of this Monte Carlo study indicate that, at a sample size of 264, coverage (proportion of results on simulated data for which the 95% confidence intervals include the true parameter value; L. K. Muthén & Muthén, 2002) and power of path coefficients (from the dummy variable to the METI dimensions) all exceed .92. Hence, we conclude that sample sizes above 264 are appropriate for Study 2. However, as greater sample sizes result in greater statistical power, we chose to recruit *at least* 264 participants from a specific list of courses, but not stop the data collection at a sample size of 264. To avoid problems through so-called optional stopping (John, Loewenstein, & Prelec, 2012), we first carried out all surveying (throughout the listed courses) before starting data analysis. The raw data sets from both studies will be published and archived via PsychData (Leibniz Institute for Psychology Information, Trier, 2018) and are also available at the corresponding Open Science Framework repository (Merk & Rosman, 2019).

Results

Sample. Following our sampling plan, we reached the intended sample size after the first course, which led to a final sample size of N = 278 ($M_{Semester}$ = 7.41, $SD_{Semester}$ = 0.30; 187 females). The proportion of participants studying least one STEM subject was 36.0%.

Measurements. We investigated the psychometric properties of the measurement instruments following our preregistered analysis plan. The main results are shown in Table 5, with additional details presented in the Reproducible Analysis Report of Study 2 (see online Supplemental Appendix 6). Overall, the factor structures of the instruments were confirmed with the exception of the treatment check (see below); reliabilities were fairly high, with all McDonald's ω values exceeding .73.

Treatment Check. In a departure from our preregistered analysis code (see online Supplemental Apeendix 5), we specified only one factor at the between-level within the MCFA (see Figure 4) of the treatment check scales, due to the poor fit of the preregistered model. This model is also theoretically plausible, as between-person scores of the treatment check can be interpreted as averages per person. This modified model yielded a very good fit, and the *y*-standardized path coefficients of the extended MIMIC model provided strong evidence for a successful treatment. For example, students who read texts containing information allegedly stemming from practitioners judged the practitioner rating scale, on average, to be more than one and a half standard deviations higher.

Research Question 1. To investigate the hypothesized "smart but evil" stereotype, we preregistered a series of models testing it at the within-person level (M1a–M1c) and at the between-person level (M2a–M2d, see the "Confirmatory Analysis Plan" section for details). The results of Models M1a to M1c can be obtained from Table 6: The regression coefficients of the dummy-coded indicator variable of the source $I^{source = pr.}$ (1 if source = practitioner, 0 otherwise) became significant in all models, and the regression weights indicated effects of moderate size in the expected direction. We thus infer that our participants exhibit a "smart but evil" stereotype at the within-person level. This stereotype manifested itself partially at the between-person level when we predicted expertise, benevolence and integrity by $I^{source = pr.}$ consecutively for each topic (M2a–M2d, see Figure 5): 9 out of 12 regression coefficients became significant, with most indicating largely moderate effect sizes.

Research Question 2. To investigate Research Question 2, we expanded M1a and M2a with topic-specific multiplism

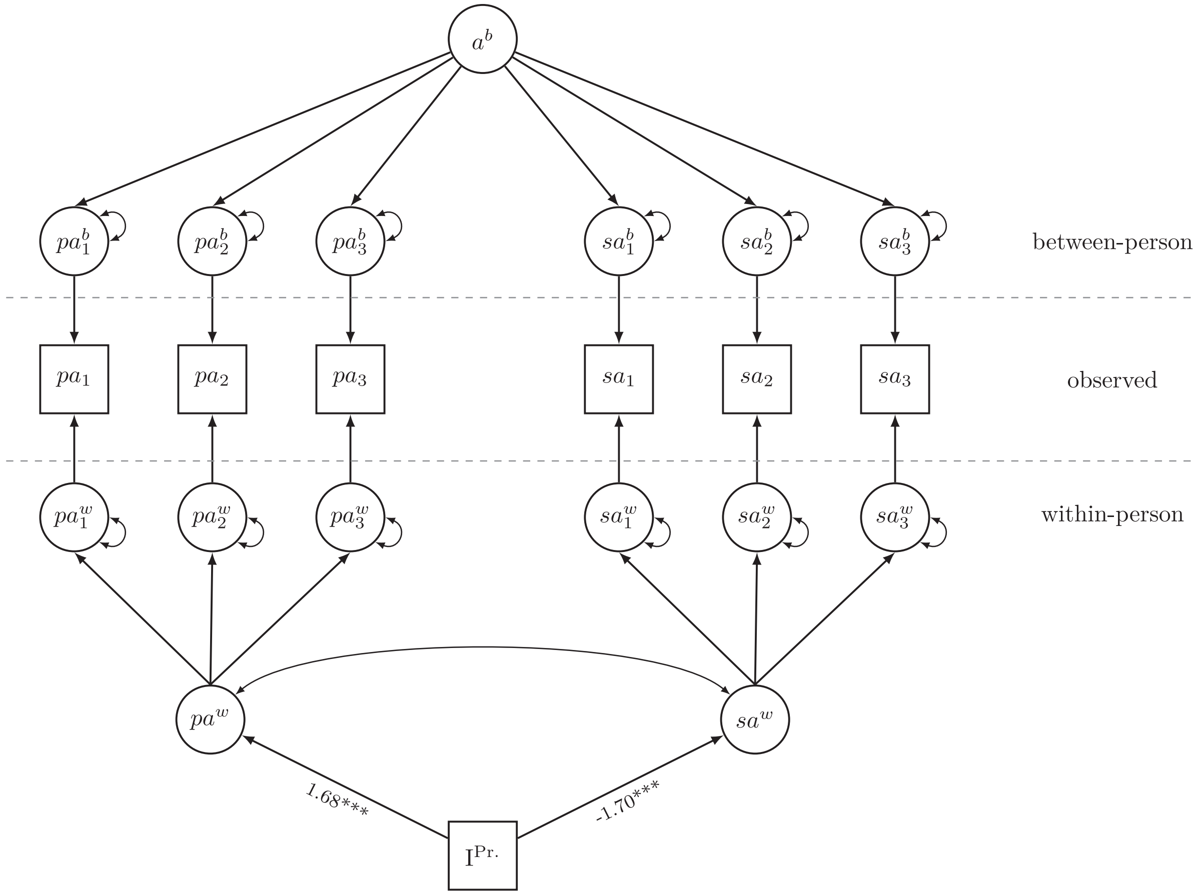

FIGURE 4. *Results of the treatment check in Study 2.*
Note. Path coefficients are *y*-standardized. a = activity; pa = practitioner activity; sa = scientist acitivity; $\mathrm{I}^{\mathrm{Pr.}}$ = dummy-coded indicator variable for the source (0 = scientist, 1 = practitioner).
***$p < .001$.

as the within-person predictor and the d-index as the between-person predictor of epistemic trust (M3a–M3c), just as we envisaged in the preregistered analysis plan. Conforming to our hypotheses, topic-specific multiplism was significantly predictive for expertise, benevolence, and integrity, revealing small to moderate effects. Contrary to our hypotheses, however, the point estimate of the regression weight of the d-index was very small and not significant.

General Discussion

In this registered report, we experimentally investigated student-teachers' epistemic trust in educational scientists compared with experts and practitioners. In one exploratory study and one strictly confirmatory and preregistered study, we found strong evidence for a "smart but evil" stereotype mainly in accordance with our hypotheses: Student-teachers judged educational scientists as having more expertise but less benevolence and less integrity than practitioners from the educational domain, whereby the between-person effects from Study 1 were larger than those from Study 2, which also were insignificant in three (out of 12) cases (Hypothesis 1). Furthermore, we found strong evidence for a negative association of topic-specific multiplism and epistemic trust (Hypothesis 2b) but more inconclusive evidence regarding a positive association of domain-specific evaluativistic beliefs and epistemic trust (Hypothesis 2a).

Apart from the benefits that arise from preregistration, the fact that we controlled for the topics and knowledge claims included in our texts underlines the robustness of these findings. However, predicting this stereotype with epistemic beliefs was only partially successful: While topic-specific multiplism was significantly related to trustworthiness in both studies, domain-specific epistemic beliefs only predicted trustworthiness in Study 1. In the paragraphs below, we first discuss the methodological strengths and weaknesses of both studies; subsequently, we suggest future directions for research and potential practical consequences for teacher education.

TABLE 6

Standardized Results of the Random Intercept Models for Research Question 2 (Study 2)

	Dependent variable											
	Expertise				Benevolence				Integrity			
	M1a		M3a		M1b		M3b		M1c		M3c	
	β (*SE*)	FMI, RIV	β (*SE*)	FMI, RIV	β (*SE*)	FMI, RIV	β (*SE*)	FMI, RIV	β (*SE*)	FMI, RIV	β (*SE*)	FMI, RIV
Intercept	**.150**	.002	.038	.002	**−.153**	.002	**−.226**	.002	**−.173**	.002	**−.246**	.002
	.049	.002	.049	.002	.050	.002	.050	.002	.048	.002	.049	.002
$\mathrm{I}^{\mathrm{source} = \mathrm{pr.}}$	**−.301**	.010	−.076	.009	**.305**	.008	**.452**	.009	**.346**	.009	**.492**	.008
	.047	.010	.045	.009	.046	.008	.047	.009	.048	.009	.049	.008
tm			**−.323**	.011			**−.211**	.008			**−.211**	.007
			.022	.011			.023	.008			.025	.006
di			.059	.023			.025	.018			.001	.012
			.044	.023			.044	.018			.042	.012
σ^2(Intercept)	.370		.400		.393		.407		.331		.346	
σ^2(Residuals)	.608		.486		.585		.534		.641		.589	
LRT *F*(*df*)	**40.07(1)**		**93.13(2)**		**63.32(1)**		**50.06(2)**		**14.64(1)**		**22.53(2)**	
RIV	.009		.019		.010		.015		.008		.009	

Note. $\mathrm{I}^{\mathrm{source} = \mathrm{pr.}}$ = dummy coded indicator (1 if source = "practicioner", 0 otherwise); tm = topic-specific multiplism; di = d-index; LRT = likelihood ratio test; RIV = relative increase in variance due to nonresponse; FMI = fraction of missing information. Boldfaced coefficients indicate $p < .05$.

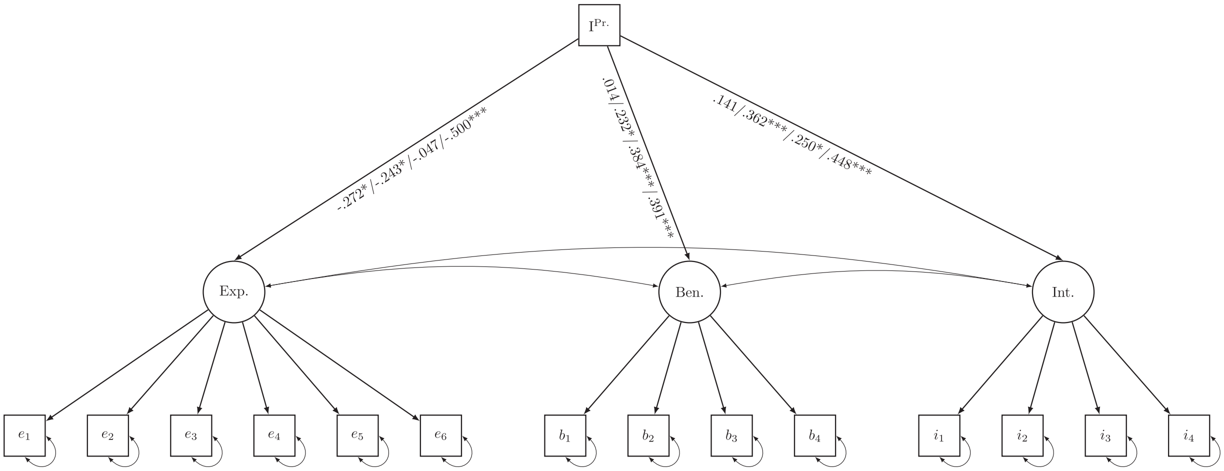

FIGURE 5. *Results of Models M2a/M2b/M2c/M2d testing the "smart but evil" hypothesis on the between-person level consecutively for each topic.*

Note. Path coefficients are *y*-standardized. Exp. = Expertise; Ben. = Benevolence; Int. = Integrity; $\mathrm{I}^{\mathrm{Pr.}}$ = dummy coded indicator variable for the source (0 = scientist, 1 = practitioner).

$^{*}p < .05$. $^{***}p < .001$.

A major strength of the present research is that we were able to replicate an interesting exploratory finding (Study 1) using a comparatively strong confirmatory approach (preregistered Study 2). In fact, preregistration has been shown to lower the likelihood of false-positive findings (Nelson, Simmons, & Simonsohn, 2018), with calls for replications becoming more pronounced in recent years (Makel & Plucker, 2014). Furthermore, investigating the "smart but evil" stereotype using an experimental design is an additional strength of our studies because carefully counterbalancing different combinations of sources and topics should increase the internal validity of our results. Finally, internal validity was also strengthened by the fact that we investigated our hypotheses using both between-person and within-person designs. Concurrently, however, external validity is limited by the fact that all participants studied at the same university.

Furthermore, it should be pointed out that 3 of 12 nonsignificant *p* values regarding the between-person effects in Study 2 were inconclusive (Amrhein, Greenland, & McShane, 2019; Dienes, 2014); it remains unclear whether they result from insufficient statistical power or the absence of effects. This points to a central weakness in our studies: Whereas presenting our participants with specific topics and assessing trustworthiness and epistemic beliefs regarding those topics allowed us to construct an internally valid study, external validity might have suffered from this approach. For example, one cannot directly conclude what would have happened if we had chosen another set of topics. Therefore, even though we chose a set of fairly typical educational topics, generalizing our findings to other topics or to the domain of educational research generally should only be done with caution. Moreover, we concede that our effect sizes were somewhat smaller than expected, which might be caused by our rather minimal manipulation (only changing certain textual cues). The effects might thus be stronger in a study with higher external validity, for example, when confronting students with actual teachers or scientists. Hence, according to Prentice and Miller (1992), even the small effects in our study might have considerable behavioral implications—but this assumption should be tested in future studies, of course.

Another inference that should be handled carefully is the results of our second research question. In fact, there are inconsistent results between the two studies (significant effects of the d-index on epistemic trustworthiness) and within Study 2 (significant effects of topic-specific multiplism, but no effects of the [domain-specific] d-index) regarding the effects of epistemic beliefs on epistemic trustworthiness. This may be due to a theoretical assumption pointed out earlier by Schraw (2001) and Bråten and Strømsø (2010), who emphasize that epistemic beliefs at different specificity levels may have the strongest impact on dependent variables that are at the same levels of specificity. This is coherent with our findings from Study 2, in which topic-specific multiplism significantly predicted topic-specific trustworthiness, whereas the domain-specific d-index did not.

In addition to the limitations mentioned above, we emphasize that the studies presented here focused on the *existence* of the "smart but evil" stereotype, not on its genesis or consequences. Both topics may be fruitfully studied in the future. The theoretical outline presented above suggests that student-teachers, on one hand, are obliged to trust the utterances of educational researchers due to the cognitive division of labor. On the other hand, their epistemic vigilance should lower the risk of being manipulated through misinformation. But why do student-teachers show higher vigilance (as shown by lower ratings of benevolence and integrity) toward educational scientists than to practitioners? This is an open question that could be investigated by referring to theories from social psychology such as intergroup relations (Brewer, 1999; Tajfel & Turner, 1986). For example, Brewer (1999) suggests that individuals usually ascribe higher trustworthiness to members of their ingroup than to those in the out-group, and student-teachers might regard actual teachers as more of an in-group than educational researchers. Another direction of future research might address the generalizability of our findings to other universities, to different academic and professional domains (beyond educational science and teaching), and to other cultural contexts. All study materials and instruments are freely available at the Open Science Framework (Merk & Rosman, 2019), and we welcome direct or conceptual replications of our studies as well as related research. In particular, it might be interesting to investigate which consequences or effects of this magnitude may show on (pre)service teacher's behavior: Will they choose other sources (academic textbook vs. blog entry by a teacher) while preparing their lessons? Will they integrate information from various sources in a different way?

With regard to the practical implications of our findings—conceding that further knowledge about the genesis of the "smart but evil" stereotype is necessary to draw strong evidence-based conclusions—several assertions can be made. First, making oneself aware of the existence of the stereotype and talking about it with students may be a first step in overcoming its problematic nature. Second, one might strive to design interventions to directly increase student-teachers' trust in educational research. In line with our deliberations on intergroup relations (see above), this might be done, for example, by referring to the method of imagined intergroup contact (e.g., Vezzali, Capozza, Stathi, & Giovannini, 2012). Third, considering the moderate impact of epistemic beliefs on epistemic trustworthiness, interventions that foster students' epistemic beliefs might also be worthwhile in this context (Kerwer & Rosman, 2018; Rosman, Mayer, Merk, & Kerwer, 2019). Finally, we would like to issue a general call for transparency in research and stronger efforts in science communication: If researchers publicly preregister their hypotheses, share their materials and data and put more effort into communicating their results modestly and in plain language, teachers (and student-teachers) might trust them more.

Acknowledgments

This research was supported in part by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63).

ORCID iD

Samuel Merk https://orcid.org/0000-0003-2594-5337

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. doi:10.1038/d41586-019-00857-9

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. doi:10.1002/mpr.329

Barzilai, S., & Eshet-Alkalai, Y. (2015). The role of epistemic perspectives in comprehension of multiple author viewpoints. *Learning and Instruction*, *36*, 86–103. doi:10.1016/j.learninstruc.2014.12.003

Bauer, J., & Prenzel, M. (2012). Science education. European teacher training reforms. *Science*, *336*, 1642–1643. doi:10.1126/science.1218387

Bråten, I., & Strømsø, H. I. (2010). When law students read multiple documents about global warming: Examining the role of topic-specific beliefs about the nature of knowledge and knowing. *Instructional Science*, *38*, 635–657. doi:10.1007/s11251-008-9091-4

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues*, *55*, 429–444.

Bromme, R., Kienhues, D., Porsch, T., Bendixen, L. D., & Feucht, F. C. (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, England: Cambridge University Press.

Bromme, R., Thomm, E., & Wolf, V. (2015). From understanding to deference: Laypersons' and medical students' views on conflicts within medicine. *International Journal of Science Education, Part B*, *5*, 68–91. doi:10.1080/21548455.2013.849017

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230–258. doi:10.1177/0049124192021002005

Buehl, M. M., & Alexander, P. A. (2001). Beliefs about academic knowledge. *Educational Psychology Review*, *13*, 385–418. doi:10.1023/A:1011917914756

Buehl, M. M., & Alexander, P. A. (2006). Examining the dual nature of epistemological beliefs. *International Journal of Educational Research*, *45*, 28–42. doi:10.1016/j.ijer.2006.08.007

Buehl, M. M., & Fives, H. (2009). Exploring teachers' beliefs about teaching knowledge: Where does it come from? Does it change? *Journal of Experimental Education*, *77*, 367–408. doi:10.3200/JEXE.77.4.367-408

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Collins, H. (2009). We cannot live by scepticism alone. *Nature*, *458*, 30. doi:10.1038/458030a

Cummings, L. (2014). The "trust" heuristic: Arguments from authority in public health. *Health Communication*, *29*, 1043–1056. doi:10.1080/10410236.2013.831685

DePuy, V., & Berger, V. W. (2014). Counterbalancing. In *Wiley StatsRef: Statistics Reference Online*. Wiley. doi:10.1002/9781118445112.stat06195

Dienes, Z. (2014, July). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 1–17. doi:10.3389/fpsyg.2014.00781

Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399–412. doi:10.1111/bjop.12046

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika*, *44*, 409–420. doi:10.1007/BF02296204

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gitlin, A., Barlow, L., Burbank, M. D., Kauchak, D., & Stevens, T. (1999). Pre-service teachers' thinking on research: Implications for inquiry oriented teacher education. *Teaching and Teacher Education*, *15*, 753–769. doi:10.1016/S0742-051X(99)00015-3

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493–498. doi:10.1111/2041-210X.12504

Greene, J. A., Azevedo, R., & Torney-Purta, J. (2008). Modeling epistemic and ontological cognition: Philosophical perspectives and methodological directions. *Educational Psychologist*, *43*, 142–160. doi:10.1080/00461520802178458

Händel, M., Artelt, C., & Weinert, S. (2013). Assessing metacognitive knowledge: Development and evaluation of a test instrument. *Journal of Educational Research Online*, *5*, 162–188.

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*, e0139309. doi:10.1371/journal.pone.0139309

Hendriks, F., Kienhues, D., & Bromme, R. (2016a). Disclose your flaws! Admission positively affects the perceived trustworthiness of an expert science blogger. *Studies in Communication Sciences*, *16*, 124–131. doi:10.1016/j.scoms.2016.10.003

Hendriks, F., Kienhues, D., & Bromme, R. (2016b). Evoking vigilance: Would you (dis)trust a scientist who discusses ethical implications of research in a science blog? *Public Understanding of Science*, *25*, 992–1008. doi:10.1177/0963662516646048

Hofer, B. K. (2001). Personal epistemology research: Implications for learning and teaching. *Educational Psychology Review*, *13*, 353–383. doi:10.1023/A:1011965830686

Hofer, B. K. (2006). Domain specificity of personal epistemology: Resolved questions, persistent issues, new models. *International Journal of Educational Research*, *45*, 85–95. doi:10.1016/j.ijer.2006.08.006

Hofer, B. K., & Bendixen, L. D. (2012). Personal epistemology: Theory, research, and future directions. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, & J. Sweller (Eds.), *APA educational psychology handbook, Vol. 1: Theories, constructs, and critical issues* (pp. 227–256). Washington, DC: American Psychological Association.

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*, 88–140. doi:10.3102/00346543067001088

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. doi:10.1080/10705519909540118

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639. doi:10.1080/01621459.1975.10482485

Keil, F. C. (2010). The feasibility of folk science. *Cognitive Science*, *34*, 826–862. doi:10.1111/j.1551-6709.2010.01108.x

Kerwer, M., & Rosman, T. (2018). Mechanisms of epistemic change: Under which circumstances does diverging information support epistemic development? *Frontiers in Psychology*, *9*, 2278. doi:10.3389/fpsyg.2018.02278

Krettenauer, T. (2005). Die Erfassung des Entwicklungsniveaus epistemologischer Überzeugungen und das Problem der Übertragbarkeit von Interviewverfahren in standardisierte Fragebogenmethoden [Measuring the developmental level of epistemological beliefs and the problem of transfering interview procedures to standardized questionnaire methods]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *37*, 69–79. doi:10.1026/0049-8637.37.2.69

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development*, *15*, 309–328.

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). Mahwah, NJ: Lawrence Erlbaum.

Landrum, A. R., Mills, C. M., & Johnston, A. M. (2013). When do children trust the expert? Benevolence information influences children's trust more than expertise. *Developmental Science*, *16*, 622–638. doi:10.1111/desc.12059

Leibniz Institute for Psychology Information, Trier. (2018). *PsychData* [Data-Sharing Platform]. Retrieved from https://www.psychdata.de

Limón, M. (2006). The domain generality-specificity of epistemological beliefs: A theoretical problem, a methodological problem or both? *International Journal of Educational Research*, *45*, 7–27. doi:10.1016/j.ijer.2006.08.002

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304–316. doi:10.3102/0013189X14545513

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*, 320–341. doi:10.1207/s15328007sem1103_2

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*, 709–734. doi:10.2307/258792

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284. doi:10.1037/1082-989X.10.3.259

Merk, S., Kelava, A., Schneider, J., Syring, M., & Bohl, T. (2017). Teacher students' epistemic beliefs about general pedagogical knowledge: Topic-, source- and context specificity. *Journal for Educational Research Online*, *9*(1), 169–189.

Merk, S., & Rosman, T. (2019). *Smart but evil? Student-teachers perception of educational researchers' epistemic trustworthiness*. doi:10.17605/OSF.IO/MYSA4

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*. Manuscript submitted for publication.

Merk, S., Rosman, T., Ruess, J., Syring, M., & Schneider, J. (2017). Pre-service teachers' perceived value of general pedagogical knowledge for practice: Relations with epistemic beliefs and source beliefs. *PLoS One*, *12*, e0184971. doi:10.1371/journal.pone.0184971

Merk, S., Schneider, J., Syring, M., & Bohl, T. (2016). Welchen Einfluss haben Quelle und Kontext auf epistemologische Überzeugungen bezüglich pädagogischen Wissens? Forschungsdaten zu einer experimentellen Untersuchung [Influence of resource and content on epistemic beliefs about general pedagogical knowledge. Research data of an experimental study]. [Data Files].Trier, Germany: Psychologisches Datenarchiv PsychData des Leibniz-Zentrums für Psychologische Information und Dokumentation ZPID. doi:10.5160/psychdata.b1tn15ep11

Muis, K. R. (2004). Personal epistemology and mathematics: A critical review and synthesis of research. *Review of Educational Research*, *74*, 317–377. doi:10.3102/00346543074003317

Muis, K. R., Bendixen, L. D., & Haerle, F. C. (2006). Domain-generality and domain-specificity in personal epistemology research: Philosophical and empirical reflections in the development of a theoretical framework. *Educational Psychology Review*, *18*, 3–54. doi:10.1007/s10648-006-9003-6

Munthe, E., & Rogne, M. (2015). Research based teacher education. *Teaching and Teacher Education*, *46*(2), 17–24. doi:10.1016/j.tate.2014.10.006

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398. doi:10.1177/0049124194022003006

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 599–620. doi:10.1207/S15328007SEM0904_8

Nadelson, L. S., & Hardy, K. K. (2015). Trust in science and scientists and the acceptance of evolution. *Evolution: Education and Outreach*, *8*, 9. doi:10.1186/s12052-015-0037-4

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534. doi:10.1146/annurev-psych-122216-011836

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . .Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. doi:10.1126/science.aab2374

Perry, W. G. (1970). *Forms of ethical and intellectual development in the college years: A scheme*. New York, NY: Holt, Rinehart & Winston.

Peter, J., Rosman, T., Mayer, A.-K., Leichner, N., & Krampen, G. (2016). Assessing epistemic sophistication by considering domain-specific absolute and multiplicistic beliefs separately. *British Journal of Educational Psychology*, *86*, 204–221. doi:10.1111/bjep.12098

Peters, R. G., Covello, V. T., & McCallum, D. B. (1997). The determinants of trust and credibility in environmental risk communication: An empirical study. *Risk Analysis*, *17*, 43–54. doi:10.1111/j.1539-6924.1997.tb00842.x

Popper, K. R. (1954). Degree of confirmation. *British Journal for the Philosophy of Science*, *5*(18), 143–149.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160–164. doi:10.1037/0033-2909.112.1.160

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Retzbach, J., Otto, L., & Maier, M. (2016). Measuring the perceived uncertainty of scientific evidence and its relationship to engagement with science. *Public Understanding of Science*, *25*, 638–655. doi:10.1177/0963662515575253

Rosman, T., Mayer, A.-K., Merk, S., & Kerwer, M. (2019). On the benefits of "doing science": Does integrative writing about scientific controversies foster epistemic beliefs? *Contemporary Educational Psychology*, *58*, 85–101. doi:10.1016/j.cedpsych.2019.02.007

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592. doi:10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Rule, D. C., & Bendixen, L. D. (2010). The integrative model of personal epistemology development: Theoretical underpinnings and implications for education. In L. D. Bendixen & F. C. Feucht (Eds.), *Personal epistemology in the classroom: Theory, research, and implications for practice* (pp. 94–123). New York, NY: Cambridge University Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi:10.1037/1082-989X.7.2.147

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*, 437–457. doi:10.1198/106186002760180608

Schraw, G. (2001). Current themes and future directions in epistemological research: A Commentary. *Educational Psychology Review*, *13*, 451–464. doi:10.1023/A:1011922015665

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80. doi:10.1177/1745691613514755

Sinatra, G. M., Kienhues, D., & Hofer, B. K. (2014). Addressing challenges to public understanding of science: Epistemic cognition, motivated reasoning, and conceptual change. *Educational Psychologist*, *49*, 123–138. doi:10.1080/00461520.2014.916216

Sjølie, E. (2014). The role of theory in teacher education: reconsidered from a student teacher perspective. *Journal of Curriculum Studies*, *46*, 729–750. doi:10.1080/00220272.2013.871754

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15–21. doi:10.3102/0013189X031007015

Stadtler, M., Scharrer, L., Macedo-Rouet, M., Rouet, J.-F., & Bromme, R. (2016). Improving vocational students' consideration of source information when deciding about science controversies. *Reading and Writing*, *29*, 705–729. doi:10.1007/s11145-016-9623-2

Tajfel, H., & Turner, J. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *The psychology of intergroup relations* (pp. 7–24). Chicago, IL: Nelson-Hall.

Thon, F. M., & Jucks, R. (2017). Believing in expertise: How authors' credentials and language use influence the credibility of online health information. *Health Communication*, *32*, 828–836. doi:10.1080/10410236.2016.1172296

Trautwein, U., & Lüdtke, O. (2007). Predicting global and topic-specific certainty beliefs: domain-specificity and the role of the academic environment. *British Journal of Educational Psychology*, *77*, 907–934. doi:10.1348/000709906X169012

Trautwein, U., & Lüdtke, O. (2009). Predicting homework motivation and homework effort in six school subjects: The role of person and family characteristics, classroom factors, and school track. *Learning and Instruction*, *19*, 243–258. doi:10.1016/j.learninstruc.2008.05.001

Trautwein, U., Lüdtke, O., & Beyer, B. (2004). Rauchen ist tödlich, Computerspiele machen aggressiv? Allgemeine und theorienspezifische epistemologische Überzeugungen bei Studierenden unterschiedlicher Fachrichtungen [Smoking kills, computer games lead to aggressive behavior? General and theory-specific epistemological beliefs in students from different fields of study]. *Zeitschrift für Pädagogische Psychologie*, *18*, 187–199. doi:10.1024/1010-0652.18.4.187

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

van der Linden, W., Bakx, A., Ros, A., Beijaard, D., & Vermeulen, M. (2012). Student teachers' development of a positive attitude towards research and research knowledge and skills. *European Journal of Teacher Education*, *35*, 401–419. doi:10.1080/02619768.2011.643401

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology: A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. doi:10.1016/j.jesp.2016.03.004

Vezzali, L., Capozza, D., Stathi, S., & Giovannini, D. (2012). Increasing outgroup trust, reducing infrahumanization, and enhancing future contact intentions via imagined intergroup contact. *Journal of Experimental Social Psychology*, *48*(1), 437–440. doi:10.1016/j.jesp.2011.09.008

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*, 632–638. doi:10.1177/1745691612463078

Wei, L. J. (1978). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association*, *73*(363), 559–563. doi:10.1080/01621459.1978.10480054

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *Journal of Experimental Education*, *80*, 26–44. doi:10.1080/00220973.2010.531299

Williams, D., & Coles, L. (2007). Evidence-based practice in teaching: An information perspective. *Journal of Documentation*, *63*, 812–835. doi:10.1108/00220410710836376

Wynne, B. (2006). Public engagement as a means of restoring public trust in science–hitting the notes, but missing the music? *Community Genetics*, *9*, 211–220. doi:10.1159/000092659

Zhao, J. H., & Schafer, J. L. (2016). pan: Multiple imputation for multivariate panel or clustered data [Computer software]. Retrieved from https://cran.r-project.org/

Authors

SAMUEL MERK is Junior Professor for Education at the University of Tübingen, Tübingen, Germany. He is interested in teacher education, epistemic beliefs, and open science.

TOM ROSMAN is a research associate at the Leibniz Institute for Psychology Information, Trier, Germany. He is interested in epistemic beliefs, information literacy, and frame of reference models.