Preregistration: Teacher’s reasons for trust and mistrust in scientific evidence: Reflecting a ‘smart but evil’ stereotype?

Tom Rosman*[1] & Samuel Merk[2]

[1]Leibniz Institute for Psychology Information (ZPID)

[2]University of Tübingen

Introduction

Several studies show that teachers tend to prefer experiential knowledge sources over scientific evidence (Cramer, 2012, 2013; Merk, Rosman, Rueß, Syring, & Schneider, 2017; Zeuch & Souvignier, 2016), making them more prone to believe in misconceptions, such as, for example, on the effects of grade retention on academic achievement. This is particularly worrying considering recent research which suggests that educational scientists might be seen as ‘smart but evil’. In fact, in two experimental studies, Merk and Rosman (2019) showed that teacher education students view educational researchers’ expertise as significantly higher than their integrity and benevolence – at least in contrast to their views on practitioners. The present research builds on this ‘smart but evil’ assumption and connects it with the question on *why* teachers trust or mistrust educational research.

In line with the work by Merk and Rosman (2019), we will investigate three potential reasons for trust and mistrust in scientific sources (adapted from the 2018 Wissenschaftsbarometer): (1) the perceived competence of educational researchers (e.g., one might mistrust researchers due to a belief that they frequently make mistakes), (2) their perceived integrity (e.g., mistrust might be rooted in the belief that researchers are somehow biased), and (3) their perceived benevolence (e.g., one might believe that researchers are only interested in their careers). Since trust and mistrust are often seen as two different concepts (e.g., Saunders & Thornhill, 2004), these three potential reasons will be assessed twice – once with regard to trust and once with regard to mistrust.

Hypotheses

In a first step, we will investigate the smart but evil stereotype with regard to teachers’ explanations of their trust and mistrust in educational science. Our underlying assumption is that if the smart but evil stereotype is also present in teachers (and not only in teacher education students, the sample which the study by Merk & Rosman, 2019 used), it will impact their explanations on why they trust or mistrust educational research. More specifically, we expect that our participants, to justify their trust in educational science, will more strongly refer to expertise-related reasons – compared to explanations focusing benevolence and integrity. In contrast, to justify their mistrust in educational science, they will more strongly refer to benevolence- and integrity-related reasons compared to expertise-related reasons. We therefore suggest the following confirmatory hypotheses:

- Concerning their reasons for trusting educational scientists, teachers will score higher on expertise-related reasons compared to benevolence-related reasons (H1a) and integrity-related reasons (H1b).
- Concerning their reasons for mistrusting educational scientists, teachers will score lower on expertise-related reasons compared to benevolence-related reasons (H1c) and integrity-related reasons (H1d).

Second, while Merk and Rosman's (2019) evidence suggests that educational researchers are subject to a smart but evil stereotype, it is not yet clear whether this stereotype is limited to educational research, or whether it applies to science in general. Tentatively, one may expect that the stereotype indeed generalizes to science as a whole. In fact, Merk and Rosman's (2019) findings are derived from analyses contrasting scientists with laypersons (both from the educational domain), with laypersons being seen as less 'smart' and less 'evil'. Hence, their findings might well be rooted in a more general 'smart but evil' stereotype that pertains to science as a whole. The pattern of results regarding the reasons for trust and mistrust in the 2018 Wissenschaftsbarometer (Wissenschaft im Dialog, 2018) further supports this assumption – even though the Wissenschaftsbarometer drew on a general population sample and not on teachers. To further investigate this, we will analyze whether the reasons for trust and mistrust suggested in Hypothesis 1 differ when teachers are requested to provide their responses with regard to scientists *in general*. We thereby expect that the pattern of reasons for trust and mistrust in science in general resembles the pattern suggested in Hypothesis 1 (i.e., the smart but evil stereotype). More specifically, we suggest the following confirmatory hypotheses:

- Concerning their reasons for trusting scientists in general, teachers will score higher on expertise-related reasons compared to benevolence-related reasons (H2a) and integrity-related reasons (H2b).
- Concerning their reasons for mistrusting scientists in general, teachers will score lower on expertise-related reasons compared to benevolence-related reasons (H2c) and integrity-related reasons (H2d).

For the hypotheses pairs H1a/H2a and H1b/H2b that are both significant, we will conduct additional exploratory analyses on a within-person level to estimate the magnitude of differences between the 'smart but evil' stereotypes regarding educational scientists and scientists in general.

Third, we will investigate, on an exploratory level and using additional data from the 2018 Wissenschaftsbarometer (Wissenschaft im Dialog, 2018), whether the general populations' explanations for their trust and mistrust in scientists also reflects the 'smart but evil' stereotype. On a descriptive level, the original 2018 Wissenschaftsbarometer supports this assumption (Wissenschaft im Dialog, 2018). However, we will check whether this still holds true when matching the Wissenschaftsbarometer sample to our teacher sample with regard to age and socioeconomic status (SES). After matching or controlling for these covariates, this confirmatory hypothesis will be tested similarly to Hypotheses 1 and 2:

- Concerning their reasons for trusting scientists in general, a general population sample matched for age and SES will score higher on expertise-related reasons compared to benevolence-related reasons (H3a) and integrity-related reasons (H3b).
- Concerning their reasons for mistrusting scientists in general, a general population sample matched for age and SES will score lower on expertise-related reasons compared to benevolence-related reasons (H3c) and integrity-related reasons (H3d).

For the hypotheses pairs H2a/H3a and H2b/H3b that are both significant, we will estimate, on a between-person level and thus using a combined dataset, the magnitude of potential differences in 'smart but evil' stereotypes between teachers and the general population (again controlling for age and SES; exploratory hypothesis; see above).

Sampling plan

Data collection procedures

A German sample of school teachers was recruited by means of a professional opinion research service (forsa). Participants were recruited by random digit dialing (mobile and landline) and data were collected through an online survey the called teachers were forwarded to. We specified the following sample properties before data collection:

- in-service teachers
- at general schools in Germany

Due to the random sampling, we expect very similar distributions between our sample and official population statistics concerning age, region (federal states of Germany), and type of school (representativeness).

While data collection was completed in November 2019, we asked forsa to withhold the dataset until January 15, 2020 and only send us a reduced dataset with $n = 36$ participants, which allowed us to write the analysis code for the preregistered hypotheses.

Target Sample Size and Sample Size Calculation

We plan to test our hypotheses using Bayes Factors for informative hypotheses (Gu, Mulder, & Hoijtink, 2018). As the investigated items are part of a larger survey with experimental parts, the sample size is already determined by the requirements of these parts ($N \approxeq 400$). Hence, we ran a Bayes Factor Design Analysis (BFDA; Schönbrodt & Wagenmakers, 2018) for this fixed N design, specifying the smallest effect size of interest to $d = .30$. As one can see from the reproducible documentation of this BFDA (see Appendix 1), the planned decision procedure results very rarely in 'inconclusive' or 'wrong' (false positive or false negative) results (see Figure 1).

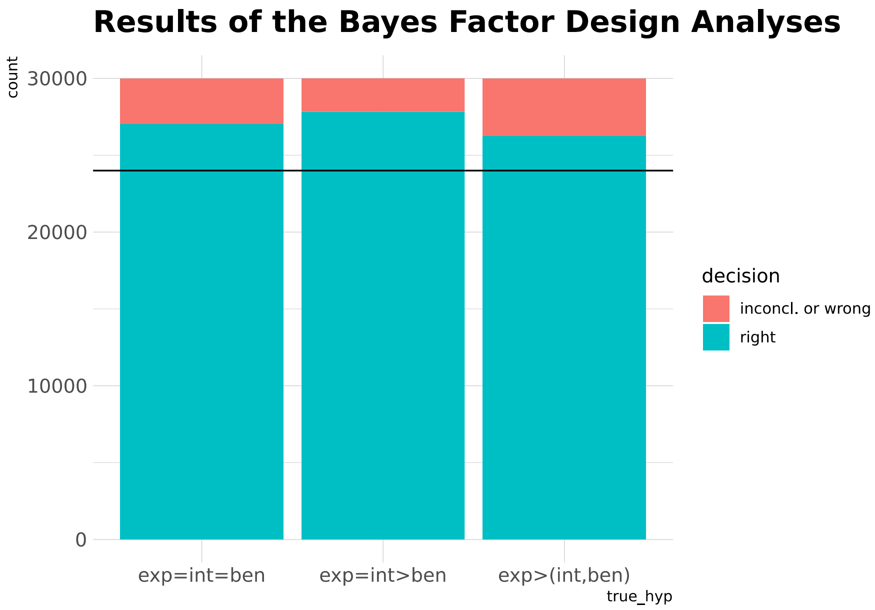

Figure 1. Results of the Bayes Factor Design Analysis (Schönbrodt & Wagenmakers, 2018). Horizontal line depicts 80% of right decisions.

Variables

Within-person factors

To investigate the exploratory part of Hypothesis 2, participants gave their responses twice – once with regard to the domain of educational science, and once with regard to science in general.

Quasiexperimental factors

The exploratory part of Hypothesis 3 will be tested by contrasting our data to the 2018 Wissenschaftsbarometer data (German general population sample; $N = 1008$; Wissenschaft im Dialog, 2018). Hence, analyses regarding Hypothesis 3 will be based on the quasiexperimental factor 'teachers' vs. 'general population'.

Experimental factors

None – even though later parts of the teacher survey drew on an experimental design, all data used for the present study was collected before any assignment to experimental groups.

Measured variables

After measuring some covariates (e.g., interest in science, general trust in science), participants were asked about their *reasons for trusting and mistrusting scientists in general*. These questions were identical to the 2018 Wissenschaftsbarometer questions (Wissenschaft im Dialog, 2018). More specifically, they were realized in a multiple-choice format with 5 response categories (range: do not agree – fully agree) and two items per reason (expertise, benevolence, integrity), of which the first one addressed reasons to trust scientists and the second one addressed reasons to *mis*trust scientists. All scales included a 'don't know' option – which will be treated as missing data and dealt with using multiple imputation if they

exceed a rate of 5 %. If the corresponding rate is lower than 5 %, we will use casewise deletion.
After responding to some more covariates, participants were asked about their *reasons for trusting and mistrusting educational science.* These questions were again identical to the Wissenschaftsbarometer questions, except for the fact that they related to educational scientists instead of scientists in general, which was realized by simply exchanging the notion 'scientists' by 'educational scientists'. Procedures regarding the 'don't know' option are the same as specified above.

Covariates
Several covariates (e.g., demographics), which are not relevant for the preregistered hypotheses outlined above, were additionally measured. All analyses regarding these covariates are exploratory.

Design plan

Study design
Even though later parts of the teacher survey drew on an experimental design, all data used for the present study was collected before assignment to any experimental groups. The study can thus be described as a simple correlational study.

Analysis plan

Statistical models for Hypotheses 1a/1b/2a/2b
Confirmatory hypotheses will be tested using Bayesian Informative Hypothesis Evaluation (Hoijtink, 2011). The following decision procedure will be used:

1) Compare approximate adjusted fractional Bayes factors for Bayesian repeated measures analysis (one within factor) for the comparisons of the following three hypotheses:

$$\mu_{Expertise} = \mu_{Integrity} = \mu_{Benevolence}$$
$$\mu_{Expertise} > (\mu_{Integrity}, \mu_{Benevolence})$$
$$\mu_{Expertise} > \mu_{Integrity} = \mu_{Benevolence}$$

2) If the data favors one hypothesis ('A') against two others ('B', 'C'), check this hypothesis against its complement ('Hc'). Else, stop the procedure with the result 'inconclusive'.

3) If the data shows evidence for the chosen hypothesis against 'Hc', stop the procedure with the result 'evidence for A in comparison to B, C and Hc'. Else, stop the procedure with the result 'inconclusive'.

The corresponding R-code can be directly deduced from the R-scripts of our BFDA (see Appendix). Within these scripts, the simulated data sets are modeled in the same way as we will analyze the empirical data. In the case of heavy violations of the model assumptions, we will switch to robust versions of these models (Bosman, 2018).

Statistical models for exploratory hypotheses

The statistical analyses for the exploratory hypotheses will not be pre-registered as they strongly depend on the data (e.g., matching is typically a data dependent, iterative process).

Inference criteria

We will judge Bayes Factors greater as 1/3 but smaller as 3 as inconclusive within our decision procedure described above.

Exclusion criteria

Missing, erroneous, or overly consistent responses: Participants may be excluded from analyses if major protocol deviations occur (e.g., selecting the same response category implausibly often). The decision to eliminate such cases will be made prior to analyzing the data. All exclusions of data will be documented in the corresponding publication(s).

References

Bosman, M. (2018). Robust Bayes factors for Bayesian ANOVA: overcoming adverse effects of non-normality (Utrecht University). Retrieved from https://informative-hypotheses.sites.uu.nl/wp-content/uploads/sites/23/2019/01/BFRobustAnova.pdf

Cramer, C. (2012). *Entwicklung von Professionalität in der Lehrerbildung: Empirische Befunde zu Eingangsbedingungen, Prozessmerkmalen und Ausbildungserfahrungen Lehramtsstudierender*. *Klinkhardt Forschung*. Bad Heilbrunn: Verlag Julius Klinkhardt.

Cramer, C. (2013). Beurteilung des bildungswissenschaftlichen Studiums durch Lehramtsstudierende in der ersten Ausbildungsphase im Längsschnitt. *Zeitschrift für Pädagogik*, *59*(1), 66–82.

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

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton: CRC Press.

Merk, S., & Rosman, T. (2019). Smart but Evil? Student-Teachers' Perception of Educational Researchers' Epistemic Trustworthiness. *AERA Open*, *5*(3), 233285841986815. https://doi.org/10.1177/2332858419868158

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*(9), e0184971. https://doi.org/10.1371/journal.pone.0184971

Saunders, M., & Thornhill, A. (2004). Trust and mistrust in organizations: An exploration using an organizational justice framework. *European Journal of Work and Organizational Psychology*, *13*(4), 493–515. https://doi.org/10.1080/13594320444000182

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

Wissenschaft im Dialog (2018). Detaillierte Ergebnisse des Wissenschaftsbarometers 2018 nach Subgruppen. Retrieved from https://www.wissenschaft-im-dialog.de/fileadmin/user_upload/Projekte/Wissenschaftsbarometer/Dokumente_18/Downloads_allgemein/Tabellenband_Wissenschaftsbarometer2018_final.pdf

Zeuch, N., & Souvignier, E. (2016). Wissenschaftliches Denken bei Lehramts- und Psychologiestudierenden. In M. Krämer, S. Preiser, & K. Brusdeylins (Eds.), *Berichte aus der Psychologie. Psychologiedidaktik und Evaluation XI* (1st ed., pp. 175–183). Herzogenrath: Shaker.

Appendix: R-Code of the BFDA

```
########################################################
#       "Bayes Factor Design Analysis (BFDA)"      ###
########################################################



library(bain)
library(psych)
library(MASS)
library(mice)
library(tidyverse)
library(hrbrthemes)



# General Approach #################
# As we plan to test our hypotheses with Bayes Factors for
# informative hypotheses [@gu2018a] we have to run a simulation
# study on our own as no dedicated R-package exists for this
# approach. Hence, we simulate 10.000 data sets for each Cohen's
# of the set $\{.0, .3,.4,.5, .6\}$, then calculate the Bayes
# Factors and note if  the results are inconclusive, provide
# evidence for the true hypothesis or provide evidence for a
# wrong hypothesis.

## Design decisions  ##################
# We did that with the following assumptions:
# As the research design involves repeated measurment we assumed
# the variance-covariance matrix
#    1 .3 .3
#   .3  1 .5
#   .3 .5  1
# as previous research has show higher correlations for the
# factors Benevolence and Integrity [@hendriks2015; @merk2019c]
#
# * We assumed the following decision procedure: After conducting
#   the study
#      1) Compare the 3 BF's for the comparisons of the three
#         hypotheses
#      2) If the data favors one hypotheses (A) against two
#         others (B,C),  check this hypotheses against Hu.
#          Else, stop the procedure withe the result
#         "inconclusive".
#      3) If the data shows evidence for the chosen hypothesis
#         against Hu, stop the procedure with the results »evidence
#         for A in comparison to B, C and Hu«.
#         Else, stop the procedure with the result "inconclusive"

## Simulation ###############################################################

# Bayesian repeated measures analysis (one within factor) #########
# Settings
sim_n <- 30000   # number of studies to simulate
true_d <- 0.3    # size of Cohen's d if mean_i != mean_j
N <- 400         # fixed sample size


# Intialize tibble with results
data_imported <- tibble(
  true_hyp = character(),
  study_iteration = integer(),
  numerator = character(),
  denominator = character(),
```

```
  BF = numeric(),
  N = numeric())

################################################################################
## Loop over study
for(study_iteration in 1:sim_n){

################################################################################
## Loop over true effects

for(true_eff in c("exp=int=ben",
                  "exp>(int,ben)",
                  "exp=int>ben")){


# Generate the data ##################################
data <- data.frame(mvrnorm(n=N,
                           mu = if(true_eff == "exp=int=ben")
                                c(0,0,0) else
                                if(true_eff == "exp>(int,ben)")
                                c(true_d,rnorm(2,0,.1)[1],rnorm(1,0,.1)[1])
                                else c(true_d, true_d, 0),
                           Sigma = matrix(c( 1, .3, .3,
                                            .3,  1, .5,
                                            .3, .5,  1),
                                          3, 3)))
names(data) <- c("exp","int","ben")

within <- lm(cbind(exp,int,ben)~1, data=data)
estimate <- coef(within)[1:3]
names(estimate) <- c("exp","int","ben")
ngroup <- nrow(data)
covmatr <- list(vcov(within))
results <- bain(estimate,"exp=int=ben;exp>(int,ben);exp=int>ben",
                n=ngroup, Sigma=covmatr,
                group_parameters=3,
                joint_parameters = 0)

sim_result <- tibble(true_hyp = true_eff,
                     study_iteration = study_iteration,
                     numerator = results$hypotheses,
                     denominator = "Hu",
                     BF = results$fit$BF[1:3],
                     N = nrow(data))%>%
  full_join(tibble(true_hyp = true_eff,
                   study_iteration = study_iteration,
                   numerator = results$hypotheses,
                   `exp=int=ben` = results$BFmatrix[, 1],
                   `exp>(int,ben)` = results$BFmatrix[, 2],
                   `exp=int>ben` = results$BFmatrix[, 3],
                   N = nrow(data))%>%
              gather(denominator, BF, `exp=int=ben`,
                     `exp>(int,ben)`, `exp=int>ben`))%>%
  filter(numerator != denominator)%>%
  filter(denominator == "Hu" |
           paste(numerator, denominator, sep = ";") %in%
             c("exp=int=ben;exp>(int,ben)",
               "exp=int=ben;exp=int>ben",
               "exp>(int,ben);exp=int>ben"))%>%
  filter(!(denominator == "Hu" & (true_hyp != numerator)))


#print(sim_result)

data_imported <- bind_rows(data_imported, sim_result)
print(study_iteration)
```

```
}
}




## Recoding right and wrong decisions if `exp=int=ben`is true
`results_exp=int=ben_true` <- tibble(study_iteration = 1:sim_n,
                                    decision = rep(NA, sim_n),
                                    true_hyp = "exp=int=ben")
for (i in 1: sim_n){

  data_curr_sim_n <- filter(data_imported, study_iteration == i
                            & true_hyp == "exp=int=ben")

  if(data_curr_sim_n%>%filter(numerator == "exp=int=ben" &
                              denominator == "Hu")%>%.$BF > 3 &
     data_curr_sim_n%>%filter(numerator == "exp=int=ben" &
                              denominator == "exp>(int,ben)")%>%.$BF > 3 &
     data_curr_sim_n%>%filter(numerator == "exp=int=ben" &
                              denominator == "exp=int>ben")%>%.$BF > 3){

    `results_exp=int=ben_true`$decision[i] <- c("right")
  }
}


## Recoding right and wrong decisions if `exp>(int,ben)` is true

`results_exp>(int,ben)_true` <- tibble(study_iteration = 1:sim_n,
                                      decision = rep(NA, sim_n),
                                      true_hyp = "exp>(int,ben)")
for (i in 1: sim_n){

  data_curr_sim_n <- filter(data_imported, study_iteration == i &
                              true_hyp == "exp>(int,ben)")

  if(data_curr_sim_n%>%filter(numerator == "exp>(int,ben)" &
                              denominator == "Hu")%>%.$BF > 3 &
     data_curr_sim_n%>%filter(numerator == "exp=int=ben" &
                              denominator == "exp>(int,ben)")%>%.$BF < 1/3 &
     data_curr_sim_n%>%filter(numerator == "exp>(int,ben)" &
                              denominator == "exp=int>ben")%>%.$BF > 3){

    `results_exp>(int,ben)_true`$decision[i] <- c("right")
  }
}


## Recoding right and wrong decisions if `exp=int>ben` is true
`results_exp=int>ben_true` <- tibble(study_iteration = 1:sim_n,
                                    decision = rep(NA, sim_n),
                                    true_hyp = "exp=int>ben")
for (i in 1: sim_n){

  data_curr_sim_n <- filter(data_imported, study_iteration == i &
                              true_hyp == "exp=int>ben")

  if(data_curr_sim_n%>%filter(numerator == "exp=int>ben" &
                              denominator == "Hu")%>%.$BF > 3 &
     data_curr_sim_n%>%filter(numerator == "exp=int=ben" &
                              denominator == "exp=int>ben")%>%.$BF < 1/3 &
     data_curr_sim_n%>%filter(numerator == "exp>(int,ben)" &
                              denominator == "exp=int>ben")%>%.$BF < 1/3){
```

```
    `results_exp=int>ben_true`$decision[i] <- c("right")
  }
}


## Displaying the results

full_join(`results_exp=int>ben_true`,
                              full_join(`results_exp>(int,ben)_true`,
                                        `results_exp=int=ben_true`))%>%
  mutate(decision = ifelse(is.na(decision), "inconcl. or wrong", "right"))%>%
  ggplot(., aes(true_hyp, fill = decision)) + geom_bar() +
  ggtitle("Results of the Bayes Factor Design Analyses") +
  geom_hline(yintercept = 0.80*30000) +
  hrbrthemes::theme_ipsum()
```