Journal of MultiDisciplinary Evaluation
Volume 14, Issue 30, 2018

ISSN 1556-8180
http://www.jmde.com

Treatment Effect Estimation Using Self-Estimated Counterfactuals Under Varying Conditions: A Meta-Analytic Exploration

Christoph Emanuel Mueller
German Research Institute for Public Administration, Institute for Regulatory Impact Assessment and Evaluation

Hansjoerg Gaus
Saarland University, Center for Evaluation

Background: Randomized controlled trials (RCTs) are frequently not an option in evaluation practice, which is why evaluators switch to non-experimental methods–such as the "counterfactual as self-estimated by program participants" (CSEPP) for estimating intervention effects. Unfortunately, no systematic attempt has been made to test under what conditions CSEPP provides valid estimates.

Purpose: As a first step in this direction, this research compared the performance of CSEPP in terms of bias when applied in various groups of participants with various levels of education, when used for assessing the effects on various outcome variables, and when employed with various question orders within the questionnaire.

Setting: N/A

Intervention: The treatment used in this research was a short educational video, in which the audience is informed about important concepts and aspects of organ donation.

Research Design: Because investigating bias in CSEPP is difficult at participant level, a series of 40 trials was conducted and bias was analyzed at trial-level. For each trial, the effect of the same treatment was estimated by CSEPP and compared with the effect estimated by a simultaneously conducted RCT. Afterwards, we analyzed differences between CSEPP and experimental results as a function of the conditions under which the single trials took place. Despite small sample sizes of the single trials, the meta-analysis was sufficiently powered to detect even small differences between CSEPP and RCT.

Data Collection and Analysis: The data was collected via online surveys on a crowdsourcing portal. For data analysis, we applied meta-analytic methods such as random-effects meta-analysis and meta-regression.

Findings: Results show that CSEPP provided accurate effect estimates, no matter under what conditions the method was applied.

Keywords: *counterfactual self-estimation; self-estimation bias; treatment effect; research design; CSEPP*

Introduction

Generally, randomized controlled trials (RCTs) are thought to serve best for the purpose of estimating causal intervention effects. Unfortunately, RCTs are not always feasible in evaluation practice–for example because evaluators are not consulted until an intervention has already started (Shadish, Cook, & Campbell, 2002) or due to budget constraints–which is why evaluators instead often apply non-experimental methods for assessing intervention effects.

A recently introduced approach, denoted as the "counterfactual as self-estimated by program participants" (CSEPP; Mueller, Gaus, & Rech, 2014; Mueller & Gaus, 2015), capitalizes on people's ability to think counterfactually (e.g., Roese & Olson, 2014) and builds on the idea that intervention participants are capable of directly estimating their counterfactual scenario, that is, the state they would have been in after an intervention without having participated. In previous studies it was found that CSEPP worked relatively well for assessing the effects of communicative interventions on various types of self-reported attitude and behavioral intention (Mueller, Gaus, & Rech, 2014; Mueller and Gaus, 2015).

Basically, the utility of CSEPP for evaluation practice depends on its scope and possible applications. Unfortunately, so far no systematic attempt has been made to investigate whether CSEPP provides reliable estimates of treatment effects under various conditions. If the method failed to do so, it would be less useful for application in evaluation practice. In any case, evaluators should be aware of potential limitations of CSEPP. Thus, as a first step in this direction, this research compared the performance of CSEPP in terms of bias when it is applied in various groups of participants with different levels of education, for assessing the effects on various outcome variables, and with various degrees fo proximity of current and self-estimated counterfactual ratings within the questionairre. We compared groups with various levels of education, and assessed the effects on various outcome variables, and with various degrees of proximity of current and self-estimated counterfactual ratings within the questionnaire.

Because investigating bias in CSEPP studies is difficult at the participant level, a series of 40 small trials was conducted and bias was analyzed at the trial level using meta-analysis. Each trial included the same treatment, an educational film about organ donation which is similar to many videos published on the web (e.g., Tian, 2010) and used by health insurance companies or health education organizations for informing people about the issue. For each of the trials, a treatment effect was estimated by CSEPP and compared with the effect estimated by a simultaneously conducted RCT. Afterwards, an analysis was made of whether differences in estimated treatment effects between CSEPP and RCT across the trials could be explained by variation in participants' level of education, variation in outcome variables, and variation in the order of the questions within the questionnaire.

This paper is organized as follows: First, we describe the conceptual background of CSEPP and present the propositions guiding this research. Subsequently, we present the data and the methods employed, describe and discuss the observed results, and conclude by considering issues of generalizability of the findings.

Conceptual Background

According to the potential outcome model (Rubin, 1979; Holland, 1986), a causal effect is the difference in an outcome variable Y between a participant after having received the treatment and the same person under the same conditions without having received it. However, only one of the two states can be observed at a given point in time. In order to deal with this problem, evaluators frequently employ control groups for approximating the state in which participants would have been without having been exposed to the treatment. This state is frequently referred to as the *counterfactual*.

In contrast to RCTs or nonequivalent comparison group designs, CSEPP relies on participants' self-estimation of the hypothetical counterfactual as an approximation of the true but non-observable counterfactual. In practice, the approach works quite simply (Mueller, Gaus, & Rech, 2014): first, participants are asked to provide information about the outcome variable Y after having participated in an intervention. Secondly, they are asked to provide information about what Y would have been like under the condition of not having participated. The difference between these current and counterfactual ratings is employed as an estimate of the treatment effect on the treated.

Therefore, similar to the application of retrospective pretest methodology (Farel, Umble, & Polhamus, 2001; Skeff, Stratos, & Bergen, 1992), using CSEPP relies solely on the self-assessments of participants after having participated in the intervention. However, CSEPP differs from retrospective pretests because participants estimate a hypothetical state in the present rather than attempting to reconstruct the past (Mueller, Gaus, & Rech, 2014). Despite this difference, in CSEPP respondents may also use *retrospective thinking* to estimate their own counterfactual by thinking back to what their state in the outcome Y was prior to the intervention.

Following Mueller, Gaus, and Rech (2014), however, participants may also employ *counterfactual thinking* (Roese & Olson, 2014; Roese, 1997) as a strategy for estimating their own counterfactual in Y. This involves a participant mentally creating possible alternatives to events that have already occurred in the past. By applying CSEPP, participants are asked directly about such an alternative event, namely their state in Y without having participated in the intervention. In answering the question, participants may mentally simulate various possible alternatives to their current outcome in Y after participation and choose the most likely alternative as an estimate of their counterfactual.

Generally, using the difference between current and self-estimated counterfactual ratings of participants in Y as an estimate for the causal intervention effect may be biased because of participants' over- or under-estimation of the true but non-observable counterfactual. Given that the counterfactual is a scenario in which participants have never actually been, it seems reasonable to assume that there is some deviation between self-estimated and true counterfactuals. This bias–which equals the difference between the true treatment effect on a participant and the treatment effect on the same person estimated by CSEPP–is denoted as *self-estimation bias* (SEB) (Mueller, Gaus, & Rech, 2014).

Research Propositions

At the moment, there is no evidence about exactly how SEB is determined or under what conditions CSEPP provides valid effect estimates. However, we do distinguish three dimensions on which varying conditions could alter the performance of CSEPP, namely characteristics of participants, attributes and focus of the intervention, and characteristics of the method used for data collection. Because resources for a research project like this are limited and the number of conditions under which CSEPP may perform better or worse is not, we selected one variable from each of the three dimensions to be further investigated with respect to their effects on the performance of CSEPP. Precisely, we assessed whether manipulating three independent variables–representing various conditions under which CSEPP is applied–leads to variation in SEB or not. In the section that follows, we present our assumptions about the relationships between the independent variables and SEB and introduce the corresponding research propositions.

Dimension 1: Individual Characteristics

First, variation in individuals' characteristics may be responsible for varying magnitudes of SEB in CSEPP studies (Mueller & Gaus, 2015). Such characteristics could be, for example, cognitive abilities, prior participation in similar interventions, prior experiences with the topic of the intervention, pre-conceptions of the treatment, or socio-economic variables.

In this study, we focus on investigating whether the validity of treatment effects estimated by CSEPP differs between groups of participants with various levels of education. We assume that participants with higher levels of education are capable of making more precise self-estimations of their counterfactual than participants with lower levels. This proposition is motivated by the fact that the level of education is positively correlated with cognitive abilities (e.g., Falch & Massih, 2010; Carlsson, Dahl, Öckert, & Rooth, 2015; Parisi et al., 2012), which means that on average, individuals with higher education levels possess stronger cognitive abilities than individuals with lower education levels.

This relationship becomes relevant in the context of self-estimating the counterfactual because according to Byrne (2005, p. 182), "people may create different counterfactual alternatives because of differences in their ability to think about possibilities of various sorts." A reason for this relationship can be found in the fact that counterfactual imagination–the mental creation of alternatives to reality–is considered to be a process of creative thinking (Byrne, 2005), which in turn depends on the level of cognitive abilities (Jauk, Benedek, Dunst, & Neubauer, 2013). Under the assumption that participants use counterfactual thinking as a strategy for crafting their counterfactual, we therefore suppose that their cognitive abilities are positively correlated with the accuracy of their self-estimations. More precisely, we assume that lower levels of cognitive abilities lead to higher SEB in CSEPP studies.

Given that participants with higher levels of education are supposed to possess stronger cognitive skills, and that stronger cognitive skills are presumed to enhance the accuracy of self-estimated counterfactuals, the first research proposition (P1) is formulated as: *The higher participants' level of education, the lower the SEB.*

To test this proposition, we applied CSEPP in two different populations of participants, one in which subjects had a higher level of education and one in which they had a lower level.

Dimension 2: Characteristics and Focus of the Intervention

The magnitude of SEB may also vary when CSEPP is conducted under conditions with various kinds of interventions, durations, subject areas, strengths of treatment effects, or types of outcome variable affected by the intervention (Mueller, Gaus, & Rech, 2014; Mueller & Gaus, 2015).

Here we focus on whether CSEPP performs equally well when assessing treatment effects on various outcome variables. More precisely, we tested whether CSEPP delivered estimates of varying accuracy when it was applied for estimating the effects of the educational film on topic-related attitudes and topic-related knowledge. The second research proposition (P2) is formulated as: *The magnitude of SEB differs between CSEPP effect estimates on topic-related attitudes and CSEPP effect estimates on topic-related knowledge*.

This proposition was inspired by hints in the study of Mueller and Gaus (2015) that CSEPP may not perform equally well in estimating treatment effects on various types of outcome variable. This may be the case for two reasons. *The first reason* has to do with the level of difficulty of self-estimating the counterfactual for a given outcome variable. For example, Mueller and Gaus (2015) found that CSEPP provided more accurate estimates when assessing treatment effects on various kinds of attitude than when estimating effects on self-reported behavior. The authors explain this finding by the fact that behavior is determined by a wide variety of factors at both individual and macro level, and that it is difficult for participants "to reflect all these determinants (and the interactions between them) and to weight the influence of the treatment" in comparison to these key drivers (Mueller and Gaus, 2015) when estimating their own counterfactual.

Based on this reasoning, it seems reasonable to presume that CSEPP performs less accurately when employed for estimating treatment effects on complex outcome variables. Because attitudes represent a more complex construct which is affected by more determinants than knowledge–which could make it more difficult to weight the influence of the treatment on attitudes in comparison to these other determinants–we assume that it might be easier for participants to estimate their counterfactual for knowledge than for attitudes.

The second reason is rooted in the facts that a given treatment may have effects of varying sizes on various outcome variables and that these varying effect sizes may lead to varied magnitudes of SEB. More specifically, we assume that the stronger the treatment effect of the intervention under study becomes, the less accurate the effect estimates provided by CSEPP are. The idea behind this assumption has to do with the so-called *contrast effect*, which is one of the mechanisms guiding counterfactual thinking. According to Roese (1997, p. 140), "contrast effects occur when a judgment is made more extreme via the juxtaposition of some anchor or standard (Sherif & Hovland, 1961)."

This reasoning also applies for CSEPP, in which the factual state in an outcome variable Y after having participated in an intervention may serve as an anchor for making a judgment about the counterfactual in Y. If the perceived counterfactual in Y differs from the factual state, a contrast effect may distort counterfactual judgments. For example, if a training measure aims to improve nurses' knowledge about a specific hygiene routine, they may underrate the level of knowledge they would have had without the training because it looks smaller in the light of the knowledge gains achieved through the training.

We suppose that with greater perceived differences between the factual and the counterfactual–that is, with increasing treatment effects–participants' judgments of the counterfactual become more extreme because of the contrast effect. Consequently, if one of the two outcome variables used in this research is more strongly affected by the promotion film than the other, different strengths of contrast effects might lead to differences in the magnitude of SEB.

Dimension 3: Characteristics of the Method Used for Data Collection

This dimension comprises attributes of the method used for data collection. For example, the performance of CSEPP may depend on whether data are collected online, in person, or by telephone. Moreover, its performance may be affected by how questions are phrased or by the order of the questions within the questionnaire. In this article, we focus on the arrangement of current and counterfactual ratings within the questionnaire and test proposition P3: *The closer the proximity of current and counterfactual ratings within the questionnaire, the larger the SEB.*

This proposition is based on the idea that the close proximity of factual and counterfactual ratings within the questionnaire increases the likelihood that participants will (consciously or unconsciously) manipulate their self-estimated counterfactual ratings. A theoretical justification of this assumption can be found in the literature on retrospective pretest methodology. For example, Nimon, Zigarmi, and Allen (2011, p. 9) note that placing current and retrospective items side-by-side and asking participants to rate the retrospective items relative to their current responses "creates a contextual effect in which participants attend to the contrast between the two ratings." As a consequence of the direct juxtaposition of current and retrospective items, participants may provide biased retrospective ratings, for example, because of implicit theories of change, impression management, or effort justification (e.g., Hill & Betz, 2005; Taylor, Russ-Eft, & Taylor, 2009). This means that participants may report a change in an outcome variable because they expected a change (even if there wasn't any effect at all), they may report biased ratings in order to present themselves in accordance with perceived norms or mores, or they may provide biased ratings in order to justify the time and effort they have spent participating. The same reasoning applies for CSEPP studies in which current and counterfactual ratings may be set in close proximity to each other.

To prevent manipulation, Nimon et al. (2011, p. 25) suggest that participants complete separate surveys for rating current and retrospective items, and they recommend the use of survey procedures that "inhibit participants from being able to refer to the posttest while completing the [retrospective pretest]." Administering separate surveys for current and counterfactual ratings would also be a potential option in CSEPP studies. Doing this, however, eliminates one of the great advantages of CSEPP in evaluation practice, namely the necessity of collecting data at one point in time only. A "milder" version of inhibiting participants from being able to refer to the current ratings while completing counterfactual ratings would be to separate current and counterfactual items *within the same questionnaire*, for example, by ensuring that the former are completed at the beginning and the latter at the end. Between the two blocks of items, participants can answer several other questions, which may provide room for various thoughts, reduce contextual effects, and prevent manipulation of counterfactual ratings.

To test whether this assumption holds and whether CSEPP performs equally well under two conditions of proximity of current and counterfactual ratings, we used two types of questionnaire in this study. In one of these, current and counterfactual ratings were completed directly one after the other at the beginning of the questionnaire; In the other, the current ratings were completed at the beginning and the counterfactual ratings at the end of the questionnaire, while several other questions were asked in between.

Method

This section describes the data and methods used for testing whether CSEPP performed equally well under various conditions of educational level, outcome variable, and question order.

Treatment

The treatment used in this research was a short educational video[i] of seven minutes' length. In this German-language video, the audience is educated about important concepts and aspects of organ donation. It provides information about various types of organ donation, explains how the system of organ donation is organized in Germany and other parts of Europe, and compares willingness to donate organs in Germany with that in other countries. The contents of the video are illustrated using the example of an animated character named Paul. Besides catchy pictures and charts, there is a narrator who explains all the important aspects of the topic. The primary objective of the video is to inform the audience about the issue of organ donation in a comprehensible way. However, it also tackles the question of the importance of organ donation and encourages recipients to get an organ donor card.

Research Design

Testing whether participants' educational level, the type of outcome variable, and the question order are correlated with SEB in CSEPP studies requires information about the magnitude of SEB, participants' educational level, the type of outcome variable, and the question order. If it is of interest whether participants' level of education affects SEB, for example, their individual values of SEB have to be correlated with their level of education.

Determining SEB at individual level, however, requires participants' individual treatment effects estimated by CSEPP to be compared with their respective true individual treatment effects, the difference between which equals SEB. Unfortunately, true individual treatment effects cannot be observed directly, which is why estimating SEB at participant level is difficult.

One way of approximating true treatment effects at the individual level is to use participants' pretest values in the outcome as estimates of their true counterfactual and to employ the difference between participants' current and pretest ratings as a benchmark for assessing the accuracy of individual effects estimated by CSEPP. The internal validity of this approach, however, may be low because of testing effects. Administering a pretest prior to the intervention may affect counterfactual self-estimations after the intervention because participants might be able to remember their pretest values and adjust their counterfactual ratings accordingly. Thus, it would not be the accuracy of self-estimated counterfactuals that was tested, but participants' ability to remember their pretest ratings. However, this problem could be solved by adding a parallel test such as would be carried out in a Solomon four-group design, for example.

We followed a different approach in this study. Another viable strategy to circumvent the problem is to estimate the *average* SEB by comparing the average treatment effect estimated by CSEPP with the average treatment effect estimated by a simultaneously conducted RCT within the same trial. The idea behind this strategy is that RCTs–given they are properly conducted–deliver unbiased estimates of true causal treatment effects and thus serve as valid benchmarks for assessing the performance of non-experimental methods (Cook, Shadish, & Wong, 2008). Consequently, the deviation of the average treatment effect estimated by CSEPP from the average treatment effect estimated by an RCT within the same trial can be employed as an estimate of SEB at trial level.

Yet, this approach is still not sufficient for testing whether participants' educational level, the type of outcome variable, and the question order affect the magnitude of SEB. This is because one single value of SEB estimated in a single trial cannot be correlated with any independent variables. An approach to solving this problem is to conduct several trials and analyze the relationships between SEB and the manipulated conditions in participants' educational level, the type of outcome variable, and the question order at trial level. We followed this strategy and conducted 40 small trials[ii], each including a comparison of the effects of the same treatment estimated by CSEPP and a simultaneously conducted RCT. In total, we therefore obtained 40 varying values of *potential SEB* (Δ), one for each trial. In manipulating the conditions in participants' educational level, the type of outcome variable, and the question order *across the 40 trials*, we were able to assess whether differences in Δ across the trials could be explained by the manipulation of the conditions under which CSEPP was applied. Hence, instead of conducting one large trial for testing the overall performance of CSEPP, we simulated how CSEPP performed under varying conditions by applying it in 40 small trials with various groups of participants, various types of outcome variable, and various questionnaire orders.

With regard to participants' educational level, a trial took the value 1 if it was conducted in a group of people with a higher level of education[iii]–people who had at least passed the *abitur*[iv]–and the value 0 if it took place in a group of participants with a lower level of education[v]–people who, at best, held a certificate from a *realschule*[vi]. Further, the value 1 in the variable "type of outcome variable" stands for trials where topic-related attitudes were the outcome variable, and the value 0 means that topic-related knowledge was employed as the outcome measure. Finally, with respect to question order, a trial received the value 1 if the current and the self-estimated counterfactual ratings were in close proximity to each other, and it received the value 0 if they were set apart within the questionnaire.

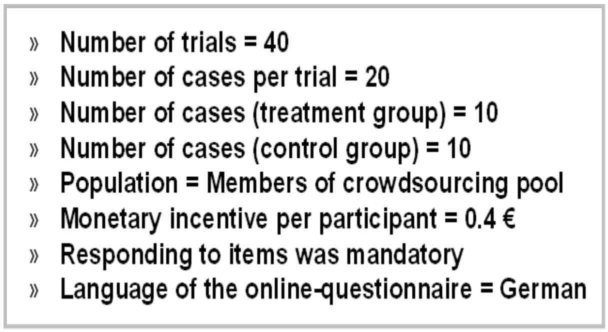
Because participants' educational level, outcome variable, and question order were dichotomous, there were 2^3 (8) possible combinations of the values of these variables. To ensure balance over the 40 trials, each of the eight combinations was assigned a lottery ticket. For each ticket, five trials were conducted. Prior to the start of each trial, a ticket was randomly drawn without replacement and the trial was implemented as specified by the ticket drawn. This procedure was repeated until five trials had been conducted for each ticket. The variable combinations of the eight tickets are presented in Figure 1.

Whereas the lottery procedure ensured proper randomization of the conditions in the type of outcome variable and the question order at trial level, we were not able to manipulate participants' educational level directly, instead letting chance decide whether a trial was to be conducted in a group of subjects with a higher or a lower level of education. Hence, educational groups may not only have varied in their level of education but also with respect to other variables that may be conducive to

SEB. In order to reduce this threat of confounding, several covariates that could vary between the educational groups were included in the analysis.

Data Collection

In total, 800 individuals were recruited for participation in the 40 trials via a German crowdsourcing portal. Each participant received a monetary incentive of 0.40 EUR (approx. 0.50 USD) for completing the whole questionnaire. Responding to all items in the questionnaire was mandatory.[vii] All the questionnaires were issued in German. Participants were assured of anonymity. The empirical work reported in this paper complied with relevant ethical standards for human subjects protections. Figure 1 provides a summary of the cornerstones of the data collection.



» Number of trials = 40
» Number of cases per trial = 20
» Number of cases (treatment group) = 10
» Number of cases (control group) = 10
» Population = Members of crowdsourcing pool
» Monetary incentive per participant = 0.4 €
» Responding to items was mandatory
» Language of the online-questionnaire = German

Figure 1. Cornerstones of Data Collection

Except for the manipulation of participants' educational level, outcome variable, and question order, the 40 trials were implemented in exactly the same manner. Each of the 40 trials was conducted with 20 participants who were randomly assigned to a treatment and a control group. Ten participants received the treatment and ten were members of the control group and did not receive the treatment. Members of the control group started the online questionnaire by rating a list of items for measuring the outcome variable specified by the ticket drawn (either topic-related attitudes or knowledge). Subsequently, they provided information on their age and gender.

Members of the treatment group started by watching the promotion film. Afterwards, they were asked about their current state in the outcome variable (depending on the ticket drawn, which was either topic-related knowledge or attitudes). Members of the treatment group in trials where the question order took the value 1 were then asked to estimate their individual counterfactual in the respective outcome variable[viii]. Subsequently, they were asked to provide information on five covariates. In contrast, treatment group members in trials where the question order took the value 0 did not estimate their own counterfactual until they had provided information on the five covariates. On the last page of the questionnaire, all members of the treatment group provided information on their age and gender.

Measures

For measuring the dependent variables, two multi-item measures were used. Topic-related attitudes were measured by nine items that had to be rated on a scale from zero (totally disagree) to six (totally agree). Similarly, topic-related knowledge was captured by ten items that were assessed on the same rating scale. For statistical analyses, the average item scores of the respective 7-point rating scales were used. The items of both measures were based on an online questionnaire by Lilie, Hübner, Mohs, and Vogel (http://sozpsy-forschung.psych.uni-halle.de/organspendefragebogen).

Five control variables were also included in the treatment group questionnaire to adjust for systematic differences between the educational groups. First, we developed a five-item construct for capturing participants' personal involvement in the topic of the intervention (e.g., Celsi & Olson, 1988) because there may be differences in this construct between the educational groups. Second, a six-item construct measured treatment sympathy, which may also be distributed differently across the educational groups. Third, we borrowed six items from Krell (2015) to measure mental effort to adjust for differences between the educational groups in regards to potential effort justification bias. Fourth, we used a scale from Musch, Brockhaus, and Bröder (2002) to measure impression management as differences in effect estimates between the educational groups may be confounded by different tendencies of socially desirable response. This scale is a German ten-item version of the impression management scale of Paulhus' 'Balanced Inventory of Desirable Responding' (Paulhus, 1992). All the items of involvement, treatment sympathy, mental effort, and impression management were rated on seven-point rating scales ranging from zero (totally disagree) to six (totally agree). The mean item scores of the 7-point rating scales were used for statistical analyses.

Finally, participants were asked whether they possessed an organ donor card. People with an organ donor card are supposed to be more experienced with the topic, and there may be differences in the proportion of donor card holders between the educational groups.

Descriptive statistics of all measures are presented in Table 1. A list of all the items showing how they were worded can be found in Appendix A.

Data Analysis

Prior to the main data analysis, we assessed whether the effect of the educational film was different for the two outcome variables attitudes and knowledge. We calculated standardized mean differences between the current ratings in the treatment and control groups (Cohen's *d*) and the respective standard errors of *d* for each of the 40 trials, and used those values to estimate pooled effect sizes by random-effects meta-analysis (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2010). The pooled effect size was d = 1.27 (95% CI [1.05, 1.49]) for trials with knowledge as the outcome, indicating a strong effect. In contrast, there was only a weak effect on attitudes with a pooled effect size of d = 0.32 (95% CI [0.12, 0.52]). Consequently, differences observed in the performance of CSEPP for estimating effects on attitudes and knowledge may indeed be a mixture of various levels of difficulty of self-estimating the counterfactual for the two measures *and* differences in the sizes of treatment effects on them.

In the section that follows, we present the seven steps of the main analysis.

Step 1. We estimated Cohen's *d* for each of the 40 trials by using both the RCT (d_{RCT}) and CSEPP (d_{CSEPP}). Effect sizes d_{RCT} for independent samples were estimated on the basis of differences in mean values between the current ratings in the treatment and control groups. Effect sizes d_{CSEPP} were estimated on the basis of differences in mean values between treatment group members' current and counterfactual ratings. Here, we also employed effect sizes for independent samples because using effect sizes for dependent samples would have led to an overestimation of d_{CSEPP} (Dunlap, Cortina, Vaslow, & Burke, 1996) and prevented comparability of d_{CSEPP} and d_{RCT}.

Step 2. To assess differences between the effects estimated by CSEPP and those estimated by the RCT, we calculated the difference between the two effect sizes (Δ) for every trial by subtracting d_{RCT} from d_{CSEPP}. Thus, Δ represents an estimate for SEB at trial level. A positive value of Δ stands for an overestimation of the RCT effect by CSEPP in a given trial, whilst a negative value of Δ stands for an underestimation. As a result of this step, we obtained 40 values of Δ, one for each trial.

Step 3. Because Δ is not a fixed value at trial level but is affected by random error within every trial, this source of uncertainty has to be taken into account in further analyses. In order to assess uncertainty within the single trials, we estimated standard errors for each value of Δ by using bootstrap resampling (Efron, 1979). In every trial, we drew 1,000 random samples from the original sample and re-estimated the effect sizes d_{RCT} and d_{CSEPP} as well as the respective differences Δ for each of the 1,000 samples. From the resulting sample distributions of Δ, we were able to derive standard errors of Δ for each of the 40 trials.

Step 4. Based on the 40 values of Δ and their respective standard errors we conducted an overall random-effects meta-analysis in order to obtain a pooled estimate of Δ. We used random-effects meta-analysis because the 40 trials were not implemented under the same conditions; that is, they were conducted in two different educational populations, with two different outcome measures, and with two different question orders in the questionnaire. For the calculations, we used the Stata module *metan* (Harris et al., 2008). A prospective power analysis, based on the formulae provided by Valentine, Pigott, and Rothstein (2010), revealed that the meta-analysis had an estimated power of .87 for detecting a small effect (Cohen's d=.35) with an assumed moderate degree of heterogeneity (τ^2=1.0) at an α-level of .05 (two-sided test).[ix]

Step 5. Because participants' educational level may be correlated with other potential determinants of Δ, differences in Δ between the two educational groups may not only be the consequence of educational status but also of other variables not controlled for by randomization. To adjust for systematic differences between educational groups in further analyses, we estimated a propensity score (Rosenbaum & Rubin, 1983) on the basis of the covariates topic involvement, mental effort, treatment sympathy, impression management, organ donor card, age, and gender. This propensity score is defined as the conditional probability of a trial being conducted in a group of participants with a higher level of education given the trial-level mean values of the covariates. For estimating the propensity score, we used logistic regression. McFadden's Pseudo R^2 was .42, suggesting that the groups did indeed differ in regard to the covariates.

Table 1
Descriptive Statistics for All Trial Groups

Ticket/ Trial	**Self-reported Knowledge (current rating in treatment group)**	**Self-reported Knowledge (counterfactual self-estimation in treatment group)**	**Self-reported Knowledge (current rating in control group)**	**Self-reported Attitudes (current rating in treatment group)**	**Self-reported Attitudes (counterfactual self-estimation in treatment group)**	**Self-reported Attitudes (current rating in control group)**	**Topic Involvement in treatment group**	**Treatment Sympathy in treatment group**	**Mental Effort in treatment group**	**Impression Mgmt. in treatment group**	**Organ Donor Card (% in treatment group)**	**Age (treatment group)**	**Age (control group)**	**Male (%in treatment group)**	**Male (% in control group)**
1/1	-	-	-	3.41 (1.05)	3.14 (0.64)	3.69 (0.96)	3.42 (1.22)	4.37 (1.11)	3.42 (1.20)	2.71 (0.65)	20	25.2 (4.44)	31.0 (13.73)	80	40
1/2	-	-	-	3.62 (0.74)	3.19 (0.95)	3.27 (1.20)	2.74 (1.47)	4.87 (0.52)	4.35 (1.09)	2.51 (1.31)	20	28.1 (6.30)	33.5 (14.27)	90	60
1/3	-	-	-	3.91 (0.79)	3.60 (0.87)	3.23 (0.88)	4.06 (0.76)	4.68 (0.72)	3.15 (1.43)	2.51 (0.92)	50	25.2 (5.69)	34.5 (12.60)	70	70
1/4	-	-	-	3.62 (0.78)	3.37 (0.78)	3.06 (0.97)	3.22 (1.54)	4.65 (0.86)	4.20 (1.16)	2.79 (1.39)	40	29.7 (11.48)	32.4 (14.39)	60	60
1/5	-	-	-	3.42 (1.72)	3.29 (1.42)	3.84 (0.76)	2.84 (2.12)	4.90 (0.72)	4.07 (0.85)	3.28 (0.83)	20	36.4 (12.55)	25.6 (9.38)	70	60
2/1	4.92 (0.55)	3.87 (1.30)	4.02 (0.88)	-	-	-	3.44 (1.12)	4.68 (0.79)	3.70 (0.98)	2.76 (1.25)	40	25.9 (7.37)	29.1 (14.37)	50	80
2/2	5.50 (0.44)	4.55 (1.30)	3.88 (1.21)	-	-	-	3.66 (1.51)	4.65 (0.38)	4.10 (1.47)	2.96 (0.75)	40	27.1 (6.71)	27.5 (7.79)	60	80
2/3	5.46 (0.51)	4.13 (1.31)	3.69 (1.17)	-	-	-	3.00 (1.74)	5.38 (0.62)	4.88 (0.66)	2.49 (0.96)	40	27.8 (8.02)	31.9 (12.81)	50	50
2/4	4.99 (0.66)	3.97 (1.01)	3.92 (1.11)	-	-	-	3.78 (1.09)	5.22 (0.72)	3.58 (1.09)	2.70 (0.74)	40	30.3 (11.95)	29.5 (13.48)	50	70
2/5	5.00 (0.52)	4.03 (0.95)	3.73 (1.07)	-	-	-	3.38 (1.54)	4.93 (0.55)	3.52 (1.30)	3.16 (0.86)	50	25.9 (9.80)	27.4 (10.21)	80	60
3/1	4.88 (1.05)	3.97 (1.41)	4.00 (1.43)	-	-	-	3.78 (1.54)	4.23 (1.35)	3.62 (1.28)	2.95 (0.75)	30	33.9 (11.91)	26.6 (7.78)	60	80
3/2	5.08 (0.56)	4.00 (0.77)	4.46 (0.91)	-	-	-	3.80 (1.56)	4.95 (0.76)	3.55 (1.69)	2.97 (0.82)	50	37.4 (16.24)	31.5 (13.57)	50	50
3/3	4.82 (1.07)	4.34 (1.09)	4.05 (1.20)	-	-	-	4.44 (1.34)	4.90 (0.87)	3.92 (1.31)	2.84 (1.07)	40	31.1 (11.76)	29.5 (8.22)	70	70
3/4	5.26 (0.48)	4.16 (0.93)	4.11 (0.72)	-	-	-	3.36 (1.14)	4.85 (0.99)	3.25 (1.50)	2.81 (1.03)	20	28.0 (11.03)	27.3 (4.00)	60	50
3/5	5.32 (0.60)	4.12 (1.04)	4.33 (1.08)	-	-	-	4.14 (1.08)	4.87 (0.67)	4.22 (1.02)	2.82 (1.08)	60	29.0 (9.04)	26.5 (9.59)	40	60
4/1	-	-	-	3.04 (0.76)	2.98 (1.01)	3.23 (1.04)	3.18 (1.91)	4.20 (1.03)	4.05 (1.52)	2.26 (1.08)	40	30.2 (8.75)	31.1 (10.34)	50	50
4/2	-	-	-	3.64 (1.12)	3.54 (0.67)	3.44 (0.93)	3.24 (1.59)	4.37 (1.16)	3.85 (0.64)	2.77 (0.76)	40	34.3 (11.74)	30.4 (6.75)	30	60
4/3	-	-	-	3.59 (1.11)	3.27 (1.00)	3.32 (0.82)	3.40 (1.29)	4.23 (1.06)	3.90 (1.44)	2.35 (0.96)	20	30.3 (16.10)	34.6 (11.17)	40	40
4/4	-	-	-	3.87 (0.92)	3.51 (0.71)	3.19 (0.95)	3.92 (1.54)	4.70 (0.83)	3.95 (1.02)	3.12 (0.52)	60	27.9 (8.77)	34.1 (9.96)	50	50
4/5	-	-	-	3.21 (1.26)	3.43 (1.17)	3.19 (0.44)	3.30 (1.44)	4.80 (0.91)	3.40 (1.32)	3.25 (0.69)	30	34.4 (11.63)	27.1 (4.09)	50	90
5/1	-	-	-	3.46 (0.94)	2.98 (1.10)	3.01 (0.75)	3.58 (1.44)	4.03 (1.35)	3.78 (1.19)	2.99 (0.77)	30	35.1 (9.28)	30.2 (8.16)	30	40
5/2	-	-	-	4.12 (0.76)	3.59 (0.69)	2.86 (0.91)	3.14 (1.35)	4.40 (1.01)	4.35 (1.27)	3.33 (1.11)	30	33.2 (13.03)	34.0 (8.27)	50	60
5/3	-	-	-	4.26 (0.94)	4.16 (1.12)	3.61 (0.71)	4.20 (1.48)	5.35 (0.78)	3.57 (0.98)	3.47 (1.20)	20	38.7 (18.64)	34.0 (8.25)	50	50

5/4	-	-	-	3.87 (0.85)	2.86 (0.63)	3.30 (0.82)	3.38 (1.52)	4.73 (0.88)	3.03 (1.15)	2.88 (0.91)	50	30.2 (10.46)	32.2 (10.35)	40	40
5/5	-	-	-	4.36 (0.83)	3.83 (1.08)	3.69 (0.88)	3.90 (2.21)	4.62 (1.54)	3.08 (1.51)	3.55 (0.97)	50	33.5 (10.43)	30.3 (4.92)	40	60
6/1	4.68 (1.91)	3.72 (1.48)	3.11 (1.31)	-	-	-	3.74 (1.43)	4.97 (0.87)	3.88 (1.18)	2.61 (0.78)	30	39.0 (12.21)	32.0 (10.79)	70	70
6/2	5.12 (0.79)	3.95 (1.34)	3.38 (1.18)	-	-	-	3.28 (1.85)	4.62 (1.53)	3.40 (1.30)	3.07 (1.14)	20	37.9 (15.52)	36.0 (12.44)	70	80
6/3	5.01 (1.02)	3.69 (1.58)	3.81 (0.99)	-	-	-	2.76 (1.65)	4.85 (1.12)	3.48 (1.27)	3.11 (0.99)	20	36.0 (14.19)	39.4 (17.46)	40	80
6/4	4.51 (1.10)	3.77 (1.39)	4.04 (1.22)	-	-	-	2.74 (1.67)	4.43 (1.19)	3.43 (1.01)	3.05 (0.84)	20	37.6 (15.77)	32.8 (14.54)	60	50
6/5	5.38 (0.58)	4.02 (0.82)	3.56 (0.92)	-	-	-	3.22 (1.95)	5.25 (0.58)	3.22 (1.38)	2.91 (0.64)	20	31.6 (10.60)	31.7 (14.35)	80	60
7/1	-	-	-	3.92 (0.93)	3.90 (0.97)	3.37 (1.08)	3.44 (1.68)	5.38 (0.69)	3.77 (1.60)	3.24 (0.93)	20	36.4 (13.26)	33.6 (8.82)	70	40
7/2	-	-	-	3.77 (1.01)	3.50 (0.92)	3.57 (0.66)	3.50 (1.64)	4.95 (0.90)	3.52 (1.44)	2.68 (1.10)	30	33.5 (12.77)	29.3 (10.63)	50	60
7/3	-	-	-	3.48 (0.72)	3.28 (0.80)	3.87 (0.94)	3.34 (1.48)	4.80 (0.98)	3.43 (1.05)	3.16 (0.98)	50	35.8 (13.51)	37.7 (18.05)	80	50
7/4	-	-	-	3.22 (1.26)	2.78 (1.22)	3.50 (1.23)	3.08 (1.87)	4.58 (1.16)	3.80 (1.20)	3.00 (0.65)	30	30.1 (9.24)	34.6 (13.32)	60	40
7/5	-	-	-	3.91 (0.97)	3.59 (0.84)	3.74 (1.23)	3.16 (1.24)	4.72 (1.17)	3.32 (1.30)	3.33 (1.05)	20	35.1 (13.18)	37.3 (10.70)	80	30
8/1	5.12 (0.85)	4.10 (1.27)	3.15 (0.79)	-	-	-	3.52 (1.32)	5.07 (0.77)	4.52 (0.87)	3.05 (1.35)	10	35.0 (15.61)	34.3 (13.18)	70	50
8/2	4.66 (1.02)	3.08 (1.81)	3.53 (1.04)	-	-	-	2.92 (1.53)	4.45 (1.26)	3.62 (1.07)	3.36 (1.16)	20	32.9 (14.98)	28.5 (11.02)	30	70
8/3	4.80 (1.24)	2.96 (1.48)	3.47 (1.00)	-	-	-	3.18 (1.68)	5.28 (0.99)	4.33 (1.20)	2.75 (1.30)	20	33.4 (9.45)	43.1 (12.38)	40	20
8/4	5.11 (0.48)	4.45 (0.65)	3.26 (0.92)	-	-	-	3.26 (1.02)	5.05 (0.68)	4.37 (0.88)	3.24 (1.19)	10	35.0 (10.03)	31.4 (8.15)	50	80
8/5	4.37 (1.74)	4.38 (1.32)	3.38 (1.14)	-	-	-	3.80 (1.61)	5.22 (1.08)	4.05 (1.39)	2.63 (1.20)	60	35.0 (11.61)	34.5 (14.03)	60	90
α	.90	.88	.83	.73	.71	.72	.88	.81	.75	.68	-	-	-	-	-

Note. Ticket/Trial = ticket number/trial number. Ticket 1: question order = 1, type of outcome variable = 1, educational level = 1. Ticket 2: question order = 1, type of outcome variable = 0, educational level = 1. Ticket 3: question order = 0, type of outcome variable = 0, educational level = 1. Ticket 4: question order = 0, type of outcome variable = 1, educational level = 1. Ticket 5: question order = 1, type of outcome variable = 1, educational level = 0. Ticket 6: question order = 1, type of outcome variable = 0, educational level = 0. Ticket 7: question order = 0, type of outcome variable = 1, educational level = 0. Ticket 8: question order = 0, type of outcome variable = 0, educational level = 0. α = Cronbach's alpha. Except Organ Donor Card and Male, values represent arithmetic means (standard deviations are in parentheses).

Step 6. We conducted random-effects meta-regression in order to gain deeper insights into the effects of participants' educational level, the type of outcome variable, and the question order on Δ. Participants' educational level, the type of outcome variable, and the question order were included in the regression equation as three binary predictors whereas Δ was the dependent variable. Further, we also included the estimated propensity score as a covariate in the model to adjust for systematic differences between the educational groups. For conducting meta-regression, we used the Stata module *metareg* (Harbord & Higgins, 2008).

Step 7. In the last step, we examined the predicted values of Δ in order to assess whether certain combinations of participants' educational level, the type of outcome variable, and the question order led to significant over- or underestimations of the approximated true effects when CSEPP was used.

Results

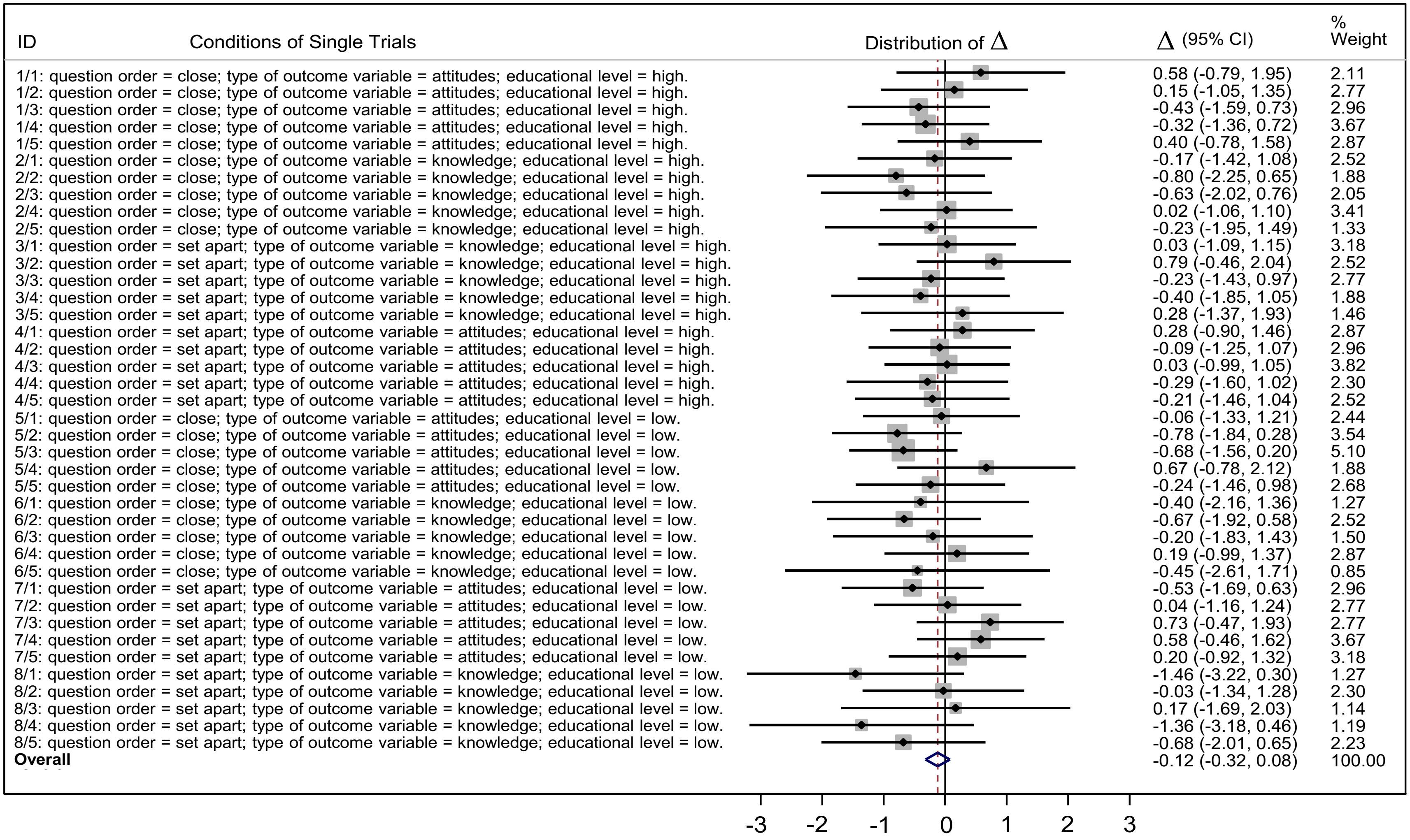
The results of the overall meta-analysis (*Step 4*) are presented in a forest plot (see Figure 2). The bold vertical line in figure 2 represents the null effect, which means that there is no difference between the effect sizes estimated by CSEPP and the simultaneously conducted RCT. The numbers presented below the horizontal line at the bottom of the plot represent standardized effect size differences in terms of Cohen's *d*. Within the plot region, each horizontal line (including small dots and boxes) represents a separate trial being analyzed by the meta-analysis. Each of these trials consists of three components: point estimates of Δ represented by the small black dots, grey boxes representing the weight each trial is given by random effects meta-analysis, and a bold horizontal line representing the 95% confidence interval of the respective point estimate, with each end of the line representing the boundaries of the confidence interval. Further, the diamond in the last row of the plot region represents the confidence interval when all the individual trials are combined. Finally, the dashed vertical line represents the point estimate of the combined individual trials.

As can be seen, there is some variation between the single trials with respect to the values of Δ. Yet, the statistical estimates suggest that this variation is almost completely determined by random noise because the estimated between-trial variation is $\tau^2 < 0.01$ and the estimated variation in Δ attributable to heterogeneity is $I^2 < 0.01\%$, $\chi^2 = 21.69$ (df 39), $p = .989$. Moreover, the pooled difference of Δ = -0.12 (95% CI [-0.32, 0.08]) is very small and does not differ significantly from zero.

Taken together, these results provide two important insights. First, the small value of the pooled Δ indicates that overall, using CSEPP for estimating treatment effects did not lead to systematic over- or underestimation when compared with an RCT. Second, the estimated between-trial variance close to zero suggests that values of Δ do not systematically vary between trials, despite the fact that these were conducted within various educational groups, with various outcome variables, and with various question orders in the questionnaire.

To gain more detailed insights into the effects of participants' educational level, the type of outcome variable, and the question order on Δ, an examination of the findings of the meta-regression (*Step 6*) is required. The estimation results support the assumption that none of the independent variables had an effect on Δ.[x] The joint test for all predictors was not significant, F (4, 35) = 0.64, $p = .637$. The same is true for the individual effects. The predictors' coefficients were $b = 0.20$ (95% CI [-0.37, 0.76]) for participants' educational level, $b = 0.21$ (95% CI [-0.21, 0.63]) for the type of outcome variable, $b = -0.20$ (95% CI [-0.61, 0.21]) for the question order, and $b = -0.07$ (95% CI [-0.87, 0.74]) for the propensity score. The intercept was $a = -0.22$ (95% CI [-0.71, 0.27]).

Finally, Figure 3 presents the predicted values of Δ for the eight tickets (*Step 7*), represented by the small diamonds in the plot region. The corresponding horizontal lines represent the respective 95% confidence intervals of the predicted values of Δ for each ticket, with each end of the line representing the boundaries of the confidence interval. In terms of effect sizes, four of the predicted values of Δ are below |0.2|, suggesting very small deviations of CSEPP estimates from RCT estimates within the respective tickets. Further, three other tickets show predicted values of Δ ranging between |0.2| and |0.3|, still representing small deviations. Only one of the fitted values (Ticket 6) exceeded the value |0.4|, indicating a small to medium deviation. Yet none of the predicted values differed significantly from zero. Consequently, there is not much evidence that specific combinations of the conditions in participants' educational level, the type of outcome variable, and the question order facilitate SEB.



ID	Conditions of Single Trials	Δ (95% CI)	% Weight
1/1:	question order = close; type of outcome variable = attitudes; educational level = high.	0.58 (-0.79, 1.95)	2.11
1/2:	question order = close; type of outcome variable = attitudes; educational level = high.	0.15 (-1.05, 1.35)	2.77
1/3:	question order = close; type of outcome variable = attitudes; educational level = high.	-0.43 (-1.59, 0.73)	2.96
1/4:	question order = close; type of outcome variable = attitudes; educational level = high.	-0.32 (-1.36, 0.72)	3.67
1/5:	question order = close; type of outcome variable = attitudes; educational level = high.	0.40 (-0.78, 1.58)	2.87
2/1:	question order = close; type of outcome variable = knowledge; educational level = high.	-0.17 (-1.42, 1.08)	2.52
2/2:	question order = close; type of outcome variable = knowledge; educational level = high.	-0.80 (-2.25, 0.65)	1.88
2/3:	question order = close; type of outcome variable = knowledge; educational level = high.	-0.63 (-2.02, 0.76)	2.05
2/4:	question order = close; type of outcome variable = knowledge; educational level = high.	0.02 (-1.06, 1.10)	3.41
2/5:	question order = close; type of outcome variable = knowledge; educational level = high.	-0.23 (-1.95, 1.49)	1.33
3/1:	question order = set apart; type of outcome variable = knowledge; educational level = high.	0.03 (-1.09, 1.15)	3.18
3/2:	question order = set apart; type of outcome variable = knowledge; educational level = high.	0.79 (-0.46, 2.04)	2.52
3/3:	question order = set apart; type of outcome variable = knowledge; educational level = high.	-0.23 (-1.43, 0.97)	2.77
3/4:	question order = set apart; type of outcome variable = knowledge; educational level = high.	-0.40 (-1.85, 1.05)	1.88
3/5:	question order = set apart; type of outcome variable = knowledge; educational level = high.	0.28 (-1.37, 1.93)	1.46
4/1:	question order = set apart; type of outcome variable = attitudes; educational level = high.	0.28 (-0.90, 1.46)	2.87
4/2:	question order = set apart; type of outcome variable = attitudes; educational level = high.	-0.09 (-1.25, 1.07)	2.96
4/3:	question order = set apart; type of outcome variable = attitudes; educational level = high.	0.03 (-0.99, 1.05)	3.82
4/4:	question order = set apart; type of outcome variable = attitudes; educational level = high.	-0.29 (-1.60, 1.02)	2.30
4/5:	question order = set apart; type of outcome variable = attitudes; educational level = high.	-0.21 (-1.46, 1.04)	2.52
5/1:	question order = close; type of outcome variable = attitudes; educational level = low.	-0.06 (-1.33, 1.21)	2.44
5/2:	question order = close; type of outcome variable = attitudes; educational level = low.	-0.78 (-1.84, 0.28)	3.54
5/3:	question order = close; type of outcome variable = attitudes; educational level = low.	-0.68 (-1.56, 0.20)	5.10
5/4:	question order = close; type of outcome variable = attitudes; educational level = low.	0.67 (-0.78, 2.12)	1.88
5/5:	question order = close; type of outcome variable = attitudes; educational level = low.	-0.24 (-1.46, 0.98)	2.68
6/1:	question order = close; type of outcome variable = knowledge; educational level = low.	-0.40 (-2.16, 1.36)	1.27
6/2:	question order = close; type of outcome variable = knowledge; educational level = low.	-0.67 (-1.92, 0.58)	2.52
6/3:	question order = close; type of outcome variable = knowledge; educational level = low.	-0.20 (-1.83, 1.43)	1.50
6/4:	question order = close; type of outcome variable = knowledge; educational level = low.	0.19 (-0.99, 1.37)	2.87
6/5:	question order = close; type of outcome variable = knowledge; educational level = low.	-0.45 (-2.61, 1.71)	0.85
7/1:	question order = set apart; type of outcome variable = attitudes; educational level = low.	-0.53 (-1.69, 0.63)	2.96
7/2:	question order = set apart; type of outcome variable = attitudes; educational level = low.	0.04 (-1.16, 1.24)	2.77
7/3:	question order = set apart; type of outcome variable = attitudes; educational level = low.	0.73 (-0.47, 1.93)	2.77
7/4:	question order = set apart; type of outcome variable = attitudes; educational level = low.	0.58 (-0.46, 1.62)	3.67
7/5:	question order = set apart; type of outcome variable = attitudes; educational level = low.	0.20 (-0.92, 1.32)	3.18
8/1:	question order = set apart; type of outcome variable = knowledge; educational level = low.	-1.46 (-3.22, 0.30)	1.27
8/2:	question order = set apart; type of outcome variable = knowledge; educational level = low.	-0.03 (-1.34, 1.28)	2.30
8/3:	question order = set apart; type of outcome variable = knowledge; educational level = low.	0.17 (-1.69, 2.03)	1.14
8/4:	question order = set apart; type of outcome variable = knowledge; educational level = low.	-1.36 (-3.18, 0.46)	1.19
8/5:	question order = set apart; type of outcome variable = knowledge; educational level = low.	-0.68 (-2.01, 0.65)	2.23
Overall		-0.12 (-0.32, 0.08)	100.00

Figure 2. Distribution of Δ.

Note. ID = Ticket/trial. Δ = Difference between effect sizes of CSEPP and RCT estimation. CI = Confidence interval of Δ. % Weight = Weights estimated by random effects meta-analysis. Overall = Pooled Δ.

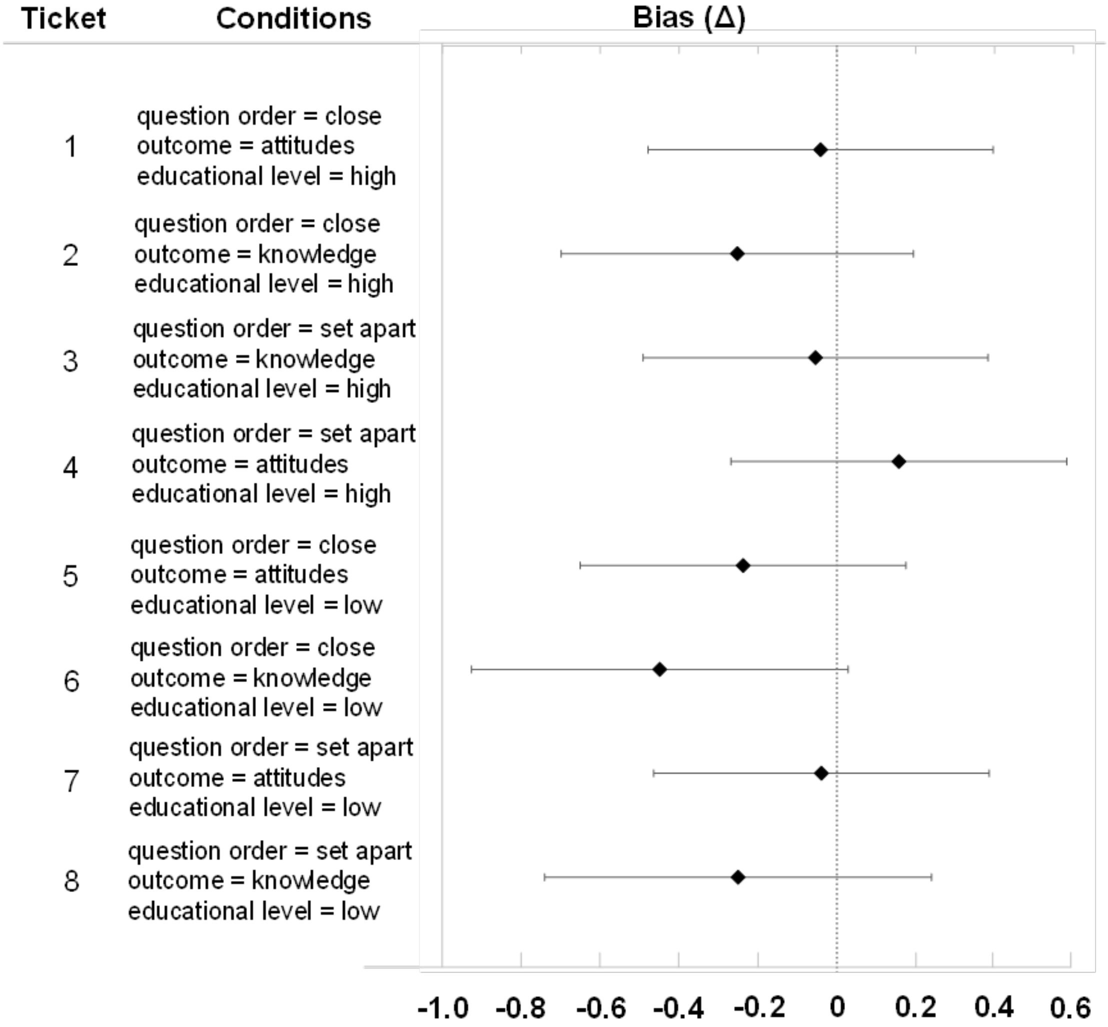


Figure 3. Predicted Value of Δ.

Note. Predicted values of Δ under control of the propensity score (held constant at its mean). Whiskers represent 95% confidence intervals of predicted values.

Discussion

The results indicate that using CSEPP did not lead to biased estimates of treatment effects when compared with a simultaneously conducted RCT. This finding confirms the results from previous studies (Mueller, Gaus, & Rech, 2014; Mueller & Gaus, 2015).

Furthermore, the estimated between-trial variation is close to zero, suggesting that values of Δ do only randomly vary. A between-trial variation of close to zero is not very common in meta-analyses because trials are usually conducted by various researchers, with various populations, and/or under various conditions. In the present study, however, all of the 40 trials were implemented in an identical manner except for the three variables manipulated. The absence of systematic between-trial variation thus leads to two insights. First, the idea of conducting every single trial under the same conditions seems to have worked out quite well. Second, if the manipulated variables had had any effect on the performance of CSEPP at all, the estimated between-trial variation would have been much higher than it was in our study. Thus, the results of our analyses clearly support the rejection of research propositions P1, P2, and P3. The important question in this regard is why.

Regarding the relationship between participants' educational level and Δ, results show that there is not much evidence for claiming an effect of participants' educational level on Δ. A first explanation for this finding is that the assumptions behind P1 are wrong, which means that there is either no correlation between the level of education and cognitive abilities, or that there is no correlation between cognitive abilities and SEB, or both. Unfortunately, it is not possible to determine which of the two assumptions may have been violated because cognitive abilities were not measured in this study.

The non-significant effect of participants' educational level on Δ may also have resulted because participants did not use counterfactual thinking for crafting their own counterfactual, but retrospective thinking instead. Given the short duration of the treatment, this is not an unlikely scenario. Thus, subjects may have tried to reconstruct the past for crafting their counterfactual instead of creating alternatives to reality. If this is the case, SEB may actually equal recall bias, a bias to which individuals are liable when they attempt to reconstruct the past, for example due to problems of remembering. Previous work has shown that recall bias is unrelated to education (Croyle et al., 2006), which may thus explain the absence of a significant effect of participants' educational level on Δ in this case.

A further explanation is related to selection bias. Because participants' educational level could not be randomized, unobserved third variables that are distributed differently between the educational groups may be responsible for confounding. Hence, although this threat to internal validity was reduced by including the propensity score in the analysis, we cannot rule out the possibility that omitted variables are responsible for the non-significant effect of participants' educational level.

Regarding the relationship between the type of outcome variable and Δ, the meta-regression did not show a significant relationship either. A first reason for this finding may be found in the absence of a contrast effect. When presenting P2, we argued that contrast effects may become stronger with increasing treatment effects, leading to increasing bias in counterfactual self-estimations. Yet this assumption may be incorrect and the presumed contrast effects may not apply.

A second potential explanation for the absence of a significant effect of the type of outcome variable on Δ in the meta-regression is that the two outcome measures are too similar in terms of difficulty for self-estimating the counterfactual. This means the assumption that it is easier for participants to estimate their counterfactual for knowledge than it is for attitudes may be false. A different explication of the absent effect may be that CSEPP is robust against varying levels of difficulty for self-estimating the counterfactual. This interpretation, however, would contradict previous findings that suggested CSEPP performed less well in cases where self-estimation of the counterfactual appears to be more difficult (Mueller & Gaus, 2015).

Finally, a third reason for the non-significant coefficient of the type of outcome variable in the meta-regression may be that varying levels of difficulty for self-estimating the counterfactual in the two outcome measures and varying strengths of treatment effects had effects on Δ in the opposite direction and thus canceled each other out. This is not unlikely given the fact that we assumed that bias due to the contrast effect is larger when estimating effects on knowledge and that bias due to the level of difficulty of self-estimating the counterfactual is larger when effects on attitudes are estimated. Unfortunately, it is not possible to separate the effects of various sizes of contrast effects and various levels of difficulty.

Another non-significant relationship was found between the question order and Δ. Here too, a first and simple explanation for this finding is that the assumptions behind P3 are incorrect. We assumed that when current and counterfactual ratings are positioned directly one after the other, participants tend to manipulate counterfactual ratings. However, the fact that participants received a monetary incentive for participation in the study may be responsible for a weakening of the tendency to manipulate because of a possible sense of duty.

Moreover, if close proximity of current and counterfactual ratings within the questionnaire promotes participants' tendency to manipulate because of implicit theories of change, effort justification, or social desirability, these sources of bias actually have to be present. However, this may not have been the case in this research. The distorting effect of implicit theories of change, for example, becomes problematic only if participants expect to change because of an intervention. Because the educational video was a low-threshold intervention of short duration, participants may, however, not have expected to change with regard to knowledge or attitudes at all. Moreover, because watching a short video is not very tiring, the tendency to justify the effort spent on watching the video is presumably low. Consequently, effort justification bias was probably not a big issue. Finally, because social desirability bias is usually low in anonymous interview situations (e.g., Tourangeau, Rips, & Rasinski, 2000), it may not have fostered manipulations of counterfactual ratings when current and counterfactual ratings were requested directly one after the other.

These explanations are supported by the results of the meta-regression. Under control of participants' educational level, the type of outcome variable, and the propensity score, the predicted value of Δ was -0.07 for trials in which the current and counterfactual ratings were placed side by side and -0.27 for trials where the ratings were separated within the questionnaire. This means that CSEPP underestimated the magnitude of treatment effects, no matter whether current and counterfactual ratings were completed directly one after the other or not. Because implicit theories of change, effort justification, and social desirability predominantly lead to the overestimation of treatment effects (e.g., Hill & Betz, 2005), it seems that these factors did not have an impact in this research.

Although CSEPP provided relatively accurate estimates regardless of the condition under which it was applied, it should be noted that CSEPP might have performed differently if the independent variables used in this study had been manipulated in a different way. As regards P1, for example, there might have been an effect of participants' educational level on Δ if the two groups had consisted of groups of academics and individuals holding a certificate of secondary education only, because of a greater educational discrepancy. Moreover, if one of the outcome variables used had measured behavior, CSEPP might have provided more varying estimates when used for estimating effects on various types of outcome variable. Finally, if current and counterfactual ratings had been completed in two separate surveys (two weeks apart, for example) we might have found effects of the question order on Δ.

Consequently, the fact that the independent variables did not have significant effects in this particular research project does not imply that they would not have had effects with other specifications.

Conclusion and Implications for Further Research

The research reported in this article was devoted to investigating the validity of CSEPP for assessing treatment effects when it is applied in various educational groups, with various outcome variables, and with various question orders in the questionnaire. The results of a comparison between CSEPP and RCT estimates showed that CSEPP delivered relatively resilient estimates of the effects of a YouTube video about organ donation on topic-related attitudes and knowledge, no matter under what conditions it was used.

When interpreting the results, one has to be aware of the fact that they only hold under the conditions examined and cannot easily be generalized. Limitations on external validity arise from various sources. For example, we worked with non-randomly sampled participants from a crowdsourcing platform only, which is why we do not know whether CSEPP would perform differently when used in various populations, such as the actual target population of the YouTube video. Moreover, the treatment was of very short duration. It is conceivable that CSEPP would have performed less satisfactorily with longer treatments. Further, the viability of CSEPP was only tested for assessing the effects of watching a film, which is why we cannot draw any inferences about whether CSEPP would have provided comparable results if the treatment had been of a different kind, such as an in-person communicative intervention or a training measure. Also, the intervention was about the issue of organ donation and we do not know whether CSEPP would have performed differently in other topical areas.

It is also difficult to generalize the results of this research study to real-world evaluation settings. For example, we do not know whether we would have obtained similar results if participants had not been volunteers but forced to participate in the intervention. Moreover, it is difficult to assess how CSEPP would perform if it was not applied directly after the intervention but with some delay. Although Mueller and Gaus (2015) showed that using CSEPP directly after the intervention did not lead to results that were substantially different from those obtained when using it two weeks later, there are currently no instances with longer time lags. Therefore, we strongly recommend that the performance of CSEPP be tested when it is applied a considerable amount of time after the intervention. This could also be helpful for determining whether CSEPP is appropriate for estimating treatment effects that build up incrementally over time, as it is unknown whether participants are capable of taking increases of effects over time into account when estimating their own counterfactual.

Furthermore, it seems reasonable to assume that it is easier for participants to estimate their own counterfactual when treatment effects are large because there is a visible contrast to their current state after the intervention. Yet, this is not the case in the present research. If we only consider the trials with attitudes as the outcome variable in the meta-analysis, we find that the pooled deviation of CSEPP from RCTs is very small at |Δ| = .04. Because the overall effect of the educational film on attitudes was small at d = .32, we conclude that CSEPP was suitable for providing relatively accurate estimates despite the fact that the treatment effect on attitudes was small.

Finally, there is some uncertainty regarding the degree to which CSEPP is capable of providing accurate estimates of complex effects where the intervention interacts with or is moderated by other variables. Also not clear is the extent to which people can anticipate side effects. Investigating both of these issues calls for the application of controlled study designs, although testing whether CSEPP is able to deal with side effects would require a wide net to be cast for additional outcomes.

As with other tests of non-experimental designs, it is difficult to draw generalizable conclusions about the conditions under which a method provides unbiased effect estimates. In this research, we only tested the effects of three various kinds of variable. However, there are many more factors that could affect the performance of CSEPP. Therefore, we recommend that the viability of CSEPP for assessing treatment effects be further investigated with more varied settings, treatments, and populations in the future. Specifically, we recommend that CSEPP be tested in real-world evaluations outside the relatively artificial research environments in which this research and previous studies on CSEPP were conducted.

Besides aspects of external validity, there are other limitations that should be noted. One problem concerns the measurement of one of the dependent variables, namely self-reported knowledge. It is unclear whether self-reported improvements in knowledge due to watching the educational video really correspond with increases in actual knowledge. Although there is evidence that self-reported knowledge measured in the context of the evaluation of a similar education video correlates at least moderately with objective knowledge measured by a multiple choice test (e.g., Mueller, 2015), this need not be the case in the present study. Consequently, we cannot draw any inferences as to whether CSEPP would have performed differently if respondents' post-video knowledge had been measured objectively by a multiple choice quiz.

Another problematic issue may be seen in the small number of cases of the single trials. Conducting an experiment with only 20 participants may lead to serious concerns on the part of many researchers. However, our ex-ante calculations showed that the overall power of our study was sufficient for detecting even small deviations of CSEPP from RCT estimates. Nevertheless, it would be beneficial to replicate the present analyses with one large trial instead of many small trials. In this case, however, problems related to estimating bias at the individual level would have to be solved in a way that was different from that employed in the present research.

Finally, while we tested whether CSEPP was capable of estimating unbiased treatment effects, we did not focus on validity issues of the self-estimated counterfactuals. For example, we did not assess whether the values of the self-estimated counterfactuals provided by respondents were actually determined by counterfactual thinking or whether respondents used other strategies such as retrospective thinking. If the latter was the case, the CSEPP method would be virtually the same as the retrospective pretest method. As long as the treatment effects estimated were unbiased, this would not compromise the CSEPP method. However, in this case the CSEPP method would have to be considered rather as a modification of the retrospective pretest and not as a new method for estimating treatment effects. Thus, we strongly suggest that future research focuses explicitly on investigating which cognitive strategies respondents actually use for creating their own counterfactuals in CSEPP studies.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. Chichester, UK: Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods, 1, 97–111.

Byrne, R. M. J. (2005). The rational imagination: How people create alternatives to reality. Cambridge: MIT Press.

Carlsson, M., Dahl, G. B., Öckert, B., & Rooth, D.-O. (2015). The effect of schooling on cognitive skills. The Review of Economics and Statistics, 97, 533–547.

Celsi, R. L., & Olson, J. C. (1988). The role of involvement in attention and comprehension processes. Journal of Consumer Research, 15, 210–224.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management, 27, 724–750.

Croyle, R. T., Loftus, E. F., Barger, S. D., Sun, Y.-C., Hart, M., & Gettig, J. (2006). How well do people recall risk factor test results? Accuracy and bias among cholesterol screening participants. Health Psychology, 25, 425–432.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods, 1, 170–177.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7, 1–26.

Falch, T., & Massih, S. (2010). The effect of education on cognitive ability. Economic Inquiry, 49, 838–856.

Farel, A., Umble, K., & Polhamus, B. (2001). Impact of an online analytic skills course. Evaluation & the Health Professions, 24, 446–459.

Harbord, R. M., & Higgins, J. P. T. (2008). Meta-regression in Stata. The Stata Journal, 8, 493–519.

Harris, R. J., Bradburn, M. J., Deeks, J. J., Harbord, R. M., Altman, D. G., & Sterne, J. A. C. (2008). Metan: Fixed- and random-effects meta-analysis. The Stata Journal, 8, 3–28.

Hill, L. G., & Betz, D. L. (2005). Revisiting the retrospective pretest. American Journal of Evaluation, 26, 501–517.

Holland, P. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 954–960.

Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. Intelligence, 41, 212–221.

Krell, M. (2015). Evaluating an instrument to measure mental load and mental effort using Item Response Theory. Science Education Review Letters, 2015, 1–6.

Mueller, C. E. (2015). Evaluating the effectiveness of website content features using retrospective pretest methodology: An experimental test. Evaluation Review, 39, 283–307.

Mueller, C. E., & Gaus, H. (2015). Assessing the performance of the "counterfactual as self-estimated by program participants": Results from a randomized controlled trial. American Journal of Evaluation, 36, 7–24.

Mueller, C. E., Gaus, H., & Rech, J. (2014). The counterfactual self-estimation of program participants: Impact assessment without control groups or pretests. American Journal of Evaluation, 35, 8–26.

Musch, J., Brockhaus, R., & Bröder, A. (2002). An inventory for the assessment of two factors of social desirability [Ein Inventar zur Erfassung von zwei Faktoren sozialer Erwünschtheit]. Diagnostica, 48, 121–129.

Nimon, K., Zigarmi, D., & Allen, J. (2011). Measures of program effectiveness based on retrospective pretest data: Are all created equal? American Journal of Evaluation, 32, 8–28.

Parisi, J. M., Rebok, G. W., Xue, Q.-L., Fried, L. P., Seeman, T. E., Tanner, E. K., Gruenewald, T. L., Frick, K. D., & Carlsson, M. C. (2012). The role of education and intellectual activity on cognition. Journal of Aging Research, 2012.

Paulhus, D. L. (1992). Assessing self-deception and impression management in self-reports: The balanced inventory of desirable responding (Reference manual, version 6). Vancouver, BC: University of British Columbia.

Roese, N. J. (1997). Counterfactual thinking. Psychological Bulletin, 121, 133–148.

Roese, N. J., & Olson, J. M. (2014). What might have been: The social psychology of counterfactual thinking. New York, NY: Psychology Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66, 688–701.

Shadish, W. R., Cook, T., & Campbell, D. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Sherif, M., & Hovland, C. I. (1961). Social judgment: Assimilation and contrast effects in communication and attitude change. New Haven, CT: Yale University Press.

Skeff, K. M., Stratos, G. A., & Bergen, M. R. (1992). Evaluation of a medical faculty development program: A comparison of traditional pre/post and retrospective pre/post self-assessment ratings. Evaluation & the Health Professions, 15, 350–366.

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Gilding the outcome by tarnishing the past: Inflationary biases in retrospective pretest. American Journal of Evaluation, 30, 31–34.

Tian, Y. (2010). Organ donation on web 2.0: Content and audience analysis of organ donation videos on YouTube. Health Communication, 25, 238–246.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. Cambridge, UK: Cambridge University Press.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. Journal of Educational and Behavioral Statistics, 35, 215–247.

Appendix A: Measures

Construct/Item

Topic-Related Attitudes
(A1) “A person who donates organs on death takes responsibility for other people.”
(A2) “Organ donation gives death some meaning.”
(A3) “I think the system we have in Germany for the allocation of donor organs is good.”
(A4) “There is a danger of organs being allocated unfairly.” (r)
(A5) “I think the way donor organs are allocated in Germany is fair and correct.”
(A6) “I would feel guilty if I refused to donate my organs on death.”
(A7) “I don’t feel any personal obligation whatsoever to donate my organs to other people when I die.” (r)
(A8) “All Germans should have an organ donor card in order to protect their family members from having to make a difficult decision when they die.”
(A9) “I don’t need an organ donor card. That’s a decision my family members can make for me when I die.” (r)

Topic-Related Knowledge
(K1) “I know what is understood by the term ‘organ donation’.”
(K2) “I know what purpose organ donations fulfill.”
(K3) “I have some idea of how the organ donation system is organized.”
(K4) “I know what a ‘post-mortem organ donation’ is.”
(K5) “I know when a person is pronounced dead in Germany so that an organ can be removed for donation.”
(K6) “I have some idea of how prepared people are to donate organs in Germany as compared with other countries.”
(K7) “I know what an organ donor card is needed for.”
(K8) “I know the difference between ‘opt-in’ and ‘opt-out’ consent systems in the context of organ donation.”
(K9) “I know how recipients of donor organs are selected in Germany.”
(K10) “I know what a ‘living organ donation’ is.”

Topic Involvement
(TI1) “The topic of organ donation is one that has been occupying me for a long time.”
(TI2) “Information on the topic of organ donation is of interest to me as a matter of basic principle.”
(TI3) “The topic of organ donation has been important to me for a long time.”
(TI4) “I have already given the topic of organ donation a good deal of consideration.”
(TI5) “The topic of organ donation doesn’t interest me.” (r)

Treatment Sympathy
(TS1) “The video was well made.”
(TS2) “The video was of high quality.”
(TS3) “The video was informative.”
(TS4) “I didn’t think much of the video.” (r)
(TS5) “The video was interesting.”
(TS6) “The video was boring.” (r)

Mental Effort
(ME1) “While watching the video I didn’t make much effort to follow the content.” (r)
(ME2) “I didn’t make any particular effort when I was watching the video.” (r)
(ME3) “I tried hard to understand everything while I was watching the video.”
(ME4) “I made a considerable mental effort while watching the video.”
(ME5) “I didn’t concentrate particularly hard while watching the video.” (r)
(ME6) “While watching the video I made an effort to follow the content.”

Impression Management
(IM1) “I sometimes tell lies if I have to.” (r)
(IM2) “There have been occasions when I have taken advantage of someone.” (r)

(IM3) “I never swear.”
(IM4) “I sometimes try to get even rather than forgive and forget.” (r)
(IM5) “I have received too much change from a salesperson without telling him or her.” (r)
(IM6) “I always declare everything at customs.”
(IM7) “I sometimes drive faster than the speed limit.” (r)
(IM8) “I have done things that I don’t tell other people about.” (r)
(IM9) “I never take things that don’t belong to me.”
(IM10) “I have taken sick-leave from work or school even though I wasn’t really sick.” (r)

Note. (r) = Item was recoded for analyses. The scale used to measure impression management was a German short form of the impression management scale of Paulhus’ Balanced Inventory of Desirable Responding (BIDR).

End Notes

[i] https://www.youtube.com/watch?v=Y0LNyK7zB88 (last accessed on December 22, 2015).

[ii] Researchers are recommended to employ at least ten trials per covariate in meta-regression models (Borenstein, Hedges, Higgins, & Rothstein, 2009). Because we included four covariates in our model, we conducted 40 trials.

[iii] 64.5% of the members of this group had passed the abitur; 35.5% held a university degree.

[iv] German school examination approximately equivalent to the American SAT exam.

[v] 1% of the members of this group left school without a degree; 4.5% were still in school; 6 % held a certificate of secondary education; 63.5% held an intermediate school certificate; 25% had completed a vocational training.

[vi] German intermediate school certificate.

[vii] Because all items were mandatory, there were no missing data.

[viii] The exact wording for asking respondents to provide self-estimations of their counterfactual was: *"Please imagine that you did not watch the film that you have just watched. How would you rate the following items then? Please be aware that you must rate these statements as you would have rated them if you had never seen the film."*

[ix] The actual power of the meta-analysis is even higher because the assumed degree of heterogeneity was substantially overstated in the prospective power calculations, as it turned out after data collection. If the estimated between-trial variance τ^2 was set to zero as suggested by the analyses, our meta-analysis had a power greater than .88 for detecting an even smaller effect (d=.25) at α=.05 (two-sided test).

[x] This is also true when the propensity score is removed from the model. Although the coefficient of participants' educational level gets somewhat smaller, the other coefficients do not change at all. Also, all confidence intervals still span the zero value.