Journal of MultiDisciplinary Evaluation
Volume 15, Issue 33, 2019

JMDE
Journal of MultiDisciplinary Evaluation

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

Retrospective Pretest and Counterfactual Self-Report: Different or Same?

Tony C.M. Lam
University of Toronto

Edgar Valencia
Centro de Medición MIDE UC, Escuela de Psicología, Pontificia Universidad Católica de Chile

Purpose: To examine discriminant validity of treatment participants' self-report of the state they would be in had they not received treatment (counterfactual); specifically, the distinction between self-report of counterfactual and self-report of preintervention state (retrospective pretest).

Setting: An education department of a large University in North America.

Intervention: Methods of self-reporting research self-efficacy with counterfactual items and with retrospective pretest items.

Research design: A randomized comparison group design with two treatments that were defined by the version of the survey used in each. In the survey for the counterfactual condition, items about research self-efficacy *without the influence* of their program of studies were included. The survey in the retrospective pretest condition contained items regarding research self-efficacy *before* participating in their program of study. The same items about research self-efficacy at the *current time* (posttest) were included in both treatment conditions.

Data collection & analysis: Participants were graduate students recruited via email who answered an online survey about research self-efficacy. These students were randomly assigned to one of the two aforementioned treatments. Responses were analyzed using a mixed 2 by 2 randomized factorial ANOVA design with self-report method (counterfactual or retrospective pretest) as the between-subjects factor and time (pre and post intervention) as the within-subjects factor.

Findings: Our findings show that counterfactual and retrospective pretest scores and treatment effects computed based on these two sets of scores are virtually identical, casting doubt on participants' ability to differentiate between a state of no treatment and a state at treatment commencement after they have received treatment.

Keywords: *retrospective pretesting; self-report; causal analysis.*

Introduction

The goal of experimentation is to isolate the impact or effect of independent variables (e.g. an intervention) on dependent variables (or outcomes). This goal is accomplished through manipulation of the independent variables in an attempt to rule out the effects of extraneous variables (confounds). The cornerstone of pinpointing treatment effect is the deployment of a counterfactual. As explained by Shadish, Cook, and Campbell:

> A counterfactual is something contrary to fact. In an experiment, we observe what *did happen* [emphasis added] when people received a treatment. The counterfactual is knowledge of *what would have happened* to those *same* people if they simultaneously had not received treatment. An *effect* is the difference between what did happen and what would have happened (2002, p. 5).

Although impossible, the idea of having the *same* individuals participate in both the treatment and control conditions *simultaneously* is ideal in experimental research; because with this arrangement, between condition comparability is assured. Since participants cannot both receive and not receive treatment at the same time, researchers create no-treatment control conditions and use participants' performance in those conditions as counterfactual (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007; Shadish et al., 2002). Although random assignment is the most effective method for assuring group equivalence, the procedure is subject to random error and procedural contaminations, which can diminish between-group-comparability and, consequently, the validity of treatment effect estimates (Cook, 2002; Dennis, 1990).

In a series of recent publications, Mueller, Gaus, and Rech (2014) and Mueller and Gaus (2015, 2018) proposed a provocative experiment design in which the *same* individuals self-reported both their post-intervention status and counterfactual estimates at post-test time. As described by Mueller and Gaus:

> First, participants are asked to provide information about the outcome variable Y after having participated in an intervention. Subsequently, they are asked to provide information about Y under the assumption of never having received the treatment. The difference between the two types of information then equals the individual treated effect on the treated. (2015, p. 8)

Mueller et al. (2014) refer to the self-report counterfactual provided by treated participants as Counterfactual Self-Estimation of Program Participants (CSEPP). Findings from their non-randomized (Mueller et al., 2014) and randomized (Mueller & Gaus, 2015) control group experiments revealed modest comparability of treatment effects based on CSEPP and self-report by control group. Mueller and Gaus concluded that results from their experiments "are encouraging for further research" (2015, p. 21). We took on their recommendation and conducted a Randomized Control Group experiment to examine the veracity of CSEPP as a measure of the no-treatment-expectation. Our research is significant because of the ground-breaking implication of CSEPP for experimental research methodology and the lack of independent research to examine the validity of CSEPP.

If CSEPP can validly mimic counterfactual by instructing participants to report counterfactual, the weakest Pre-Post-Test-Single-Group design is transformed into a design even stronger than the Randomized Control Group design. In this design, Selection Bias is no longer a validity threat given that treatment participants serve both as treatment and control subjects. In this way, the fundamental problem of causal inference – that an event cannot be present and absent simultaneously (Holland, 1986) – will finally be resolved, at least for experiments that use self-report data. Another advantage of the CSEPP design over the Randomized Control Group design is the convenience of not requiring a control group. Being able to replace control group counterfactual with self-report counterfactual would be a breakthrough in quantitative impact evaluation; the research community should scrutinize the construct validity of self-report counterfactual before fully embracing this substitution.

The Campbellion fourfold validity framework (Cook & Campbell 1979; Shadish et al. 2002), which has dominated the causal inference thinking in experimental paradigm for decades, comprises the internal, statistical conclusion, construct, and external validities. Among the various types of construct validity evidence, the relationship-to-other-variables evidences are based on the premise that a target construct should correlate with same or similar constructs but not with different or dissimilar constructs. The former is convergent validity evidence; the latter discriminant validity evidence (AERA, APA, NCME, 2014). Even though Mueller and associates found convergent validity evidence for CSEPP (similarity between CSEPP and control group counterfactual), discriminant validity evidence (dissimilarity between CSEPP and retrospective pretest self-reports) is lacking. In the influential pioneer article on convergent and discriminant validities by Campbell and Fiske (1959), the authors assert that "for the justification of novel trait measures, for the validation of test interpretation, or for the establishment of construct validity, discriminant validation as well as convergent validation is required" (p. 81). Hence, not only should CSEPP be similar to counterfactual obtained from control group (convergent validity), it should also be different (discriminant validity) if participants are asked to recall their pre-intervention status (retrospective pretest) instead of their no-treatment state or counterfactual. Mueller did not put forth the discriminant validity evidence in support of CSEPP. Our research aims to fill this void.

Retrospective Pretest and Counterfactual Self-Reports

When pre- and post-test difference is used as an estimate of treatment effect in the Pre-Post-Test-Single-Group design, three of the nine threats to internal validity (ambiguous temporal precedence, selection, and additive and interactive effects) are inconsequential; but the remaining six threats (history, maturation, regression, attrition, testing, instrumentation) are potentially present (Cook & Campbell 1979; Shadish et al. 2002). In addition, a response bias uniquely associated with pre-intervention-testing, the Response-Shift Bias (RSB), might be present (Bray, Maxwell, & Howard, 1984; Howard, 1980; Howard & Dailey, 1979; Howard, Schmeck, & Bray, 1979). RSB is the consequence of participants' lacking familiarity with or having misconceptions about treatment content before an intervention, which results in inaccurate responses to pretest surveys. Similar to the Testing and Instrumentation effects, RSB artificially inflates or deflates pretest scores, distorts pre-post change measures and subsequently, treatment effect estimates.

As a strategy to eliminate RSB, researchers and evaluators have advocated for the *retrospective pretesting* methodology, in which participants are asked to self-report both pre- and post-treatment status at the completion of an intervention (Howard, Schmeck, & Bray, 1979; Pohl, 1982). Since pre-intervention testing is no longer needed, this methodology also effectively eliminates the testing and instrumentation threats to internal validity.[1] In this way, using retrospective pretest reduces the aforementioned six internal validity threats in the Pre-Post-Test-Single-Group design to four (history, maturation, regression, and attrition).

Mueller et al. acknowledged that "the estimation of completely unbiased treatment effects, expressed as changes between pretest and posttest induced solely by treatment exposure, is…rather unlikely (White, 2010)." (2014, p. 10) Mueller et al. (2014) also noted that when pretest is substituted by retrospective pretest, treatment effect using retrospective pretest "is confronted with the same fundamental assumption that is made of pretests, namely, the stability of a unit between pretest and posttest," as well as "problems related to remembering (Howard et al, 1979) or telescoping (Rubin & Baddeley, 1989)" (2014, p. 11). Mueller et al. (2014) refer to this lack of stability across time as a result of Stability Bias (SB), which is "the sum of all

[1] Testing bias occurs if completion of a pretest affects responses to post-test, and Instrumentation bias occurs if the nature of the pretest and posttest instruments are not parallel or the testing conditions under which those instruments are administered change over time (Shaddish, Cook & Campbell, 2002, p.55).

sources of non-random errors (i.e., all internal and external factors that are not held constant over time and influence the outcome variable) occurring between pre- and post-test" (p.10). Viewed in this way, SB is analogous to threats to internal validity, and in the Retrospective-Pre-Post-Test-Single Group design, encompasses the history, maturation, regression, and attrition internal validity threats (as discussed), plus potential biases due to the delay in reporting pretest status. To improve the accuracy of treatment effect estimate based on self-report, Mueller introduces CSEPP as a replacement of retrospective pretesting in a single group design. Both CSEPP and retrospective pretest are administered along with post-testing after the completion of an intervention. The only difference is the self-report instruction; CSEPP instruction asks participants to report their state of no-treatment (counterfactual) and retrospective pretest instruction asks participants to report the state just prior to intervention (pretest).

With CSEPP, the Retrospective-Pre-Post-Test-Single-Group design is effectively transformed into a Within-Subjects-Post-Only-Control-Group design. Similar to the traditional Between-Subjects-Post-Only-Control Group design (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish et al., 2002), the within-subjects design can control for all threats to internal validity. So, when retrospective pretest self-report is replaced by CSEPP self-report, the four internal validity threats embedded in the Retrospective-Pre-Post-Test-Single-Group design should be eliminated; but measurement biases arising from the postponement of gathering pretest or CSEPP self-reports until after the intervention is completed and response biases embedded in these two types of self-report still potentially exist.[2] There is no strong reason to believe that the amount and nature of these construct related biases are different between the counterfactual and retrospective pretest thinking; however, because internal validity threats are removed in the CSEPP design, we would expect differences between CSEPP and retrospective pretest scores, as well as treatment effects estimated based on these two forms of pre-intervention assessments.

Although Mueller and associates acquired evidence of convergent validity of self-report counterfactuals, i.e., that treatment effects based on control group and CSEEP are *similar*, they did not provide the necessary discriminant validity evidence that treatment effects based on retrospective pretest and those based on CSEPP are *different.* As Mueller et al. acknowledge, "the CSEPP differs conceptually from the use of retrospective pretests because program participants estimate a hypothetical state in the present" (2014, p. 11). Without the discriminant validity evidence, the argument that CSEPP is a valid estimate of program outcome without treatment is incomplete and weak, and the assumption Mueller et al. (2014) that "program participants could estimate their own counterfactual…by mentally developing potential scenarios about how an outcome of interest (e.g., an attitude or a certain behavior) would have been like without participation in an intervention," (p. 11) cannot be substantiated.

As suggested earlier, the distinction between counterfactual and retrospective pretest self-reports is paramount. If CSEPP are just retrospective pretest estimates, it would be erroneous to give credence to CSEPP based treatment effects as a viable substitute for treatment effects derived from experiments using the Randomized Control Group design. Furthermore, believing that merely changing the instruction from reporting pretest state to reporting counterfactual state in a single group design can produce valid treatment impact estimate may encourage some researchers to lower their guard against validity threats. Hence, it is critical to strongly substantiate the validity of counterfactual generated by treatment participants, which is the goal of our research. In our search for discriminant validity evidence of CSEPP estimates, we seek the answer to this question: Is treatment impact estimate based on CSEPP different than treatment impact estimate based on retrospective pretest? In other words, are counterfactual and

[2] Mueller et al. (2014) refer to these biases as Self-Estimation Bias (SEB).

retrospective pretest self-reports different or the same?

Method

Sample

The study participant pool was graduate students from a large teacher education and research institution in North America. With approval from the ethic committee board, a recruiting email was sent to the students via the institution's mailing list. The email message asked for volunteers to participate in an online survey exploring "how education graduate students perceive their ability to do research and how to properly measure this perception" and offered the opportunity to enter a drawing for a gift card to a local bookstore after survey completion. In the informed consent, participants were assured that their answers would be anonymous and only accessible to the researchers in this study.

Due to a low number of volunteers from the first call, we recruited participants from the same pool of University students with identical method and procedure in three waves. The first wave of data collection occurred between April and July 2016, the second wave took place between April and May 2017, and the third wave took place during August 2018 producing 32, 31 and 18 participants respectively. Combining participants from the three waves created a group of 81 graduate students who completed the study.

There were neither statistically significant differences across waves nor interaction between waves and experimental condition for any of the study's dependent variables. Except for a statistically significant difference in participant's mode of study ($\chi^2(2) = 6.55$, p = 0.038) possibly due to the unbalanced number of students across waves, no significant differences in other participants' background across waves were found.

Most participants (77.78%) identified as female, with the remaining 22.22% identifying as male. Half of the respondents reported pursuing a master's degree, the other half a doctoral degree. The majority were full-time students (69.14%), and the average number of years in the program was 2.46 (*SD* = 1.70). The average number of qualitative research courses taken was 1.29 (*SD* = 1.30), slightly lower than the average number of quantitative research courses taken (*M* = 1.33, *SD* = 1.26). The proportion of students reporting at least one mixed methods research course was 41.77%. On average, students participated in 1.81 research projects (*SD* = 1.16), and 21.74% had published in peer-reviewed journals at least once. Finally, 70.37% of the students expressed a tendency towards qualitative methods and 29.63% towards quantitative methods.

Instrument

We employed an adapted version of the Clinical Research Appraisal Inventory (Mills, Caetano, & Rhea, 2014; Mullikin, Bakken, & Betz, 2007) initially developed as a self-report instrument of research skills for undergraduate and graduate medical students. The adapted scale comprises two subscales with a total of 19 items; 10 items that measure qualitative research skills, and nine that measure quantitative research skills. Examples of qualitative subscale items include the ability to gather information through observation, ability to conduct focus groups, and ability to conduct interviews. Examples of quantitative subscale items include the ability to design experimental research based on hypothesis, ability to identify independent, dependent and extraneous variables, and ability to conduct statistical analysis. Participants rated their perceived ability to conduct qualitative and quantitative research using a 7-point Likert-type response scale with only the extreme response options labelled (1 = low ability; 7 = high ability).

This study measured three types of research ability self-efficacy: Current, Past, and Counterfactual. *Current* ability refers to the ability achieved right at the end of the current term. The instruction given to students was: "The following items ask about your current abilities. For these items, please consider that current ability refers to Spring 2016/ Summer 2017/ Summer 2018." *Past* ability is the ability at the time the students entered the program (retrospective pretest). The instruction given was: "The following items ask about your perception of your past

research abilities. For these items, please consider your research ability *at the time (the month) you began your program of study* at…" The following reminder was presented just above each item: "Note that past ability refers to just before you joined [the institution]."

Counterfactual is the ability in a hypothetical situation of not having participated in the program (or CSEPP). The instruction given was: "The following items ask you to estimate what your level of ability would have been had you not joined [name of the institution]. Please be aware that these statements describe a hypothetical state. Rate these items on the assumption that you had *never joined [the institution]*, and thus *had not been exposed to* any of the courses, research, or other experiences garnered through your involvement with [the institution]". The following reminder was presented just above each item: "Rate these items on the assumption that you had never joined [the institution]." These instructions closely follow those provided by Mueller et al. (2014).

Two forms of the instrument were developed to measure research self-efficacy: (1) Retrospective Pretest (RP) and (2) Counterfactual or CSEPP. Both forms contain the same scales measuring qualitative and quantitative research skills and questions collecting background information. In the RP form, participants were instructed to report their Current and Past research abilities. In the CSEPP form, they were instructed to report their Current and Counterfactual research abilities. Item arrangement was the same under the Past and Counterfactual ability instructions, but it differed from the Current ability instruction.

Both forms produced three research ability composite scores: current qualitative, current quantitative, and current overall. The RP form also produced a past qualitative, past quantitative, and past overall research ability score; while the CSEPP form also produced a counterfactual qualitative, counterfactual quantitative, and counterfactual overall research ability score. The twelve Cronbach's alphas of item responses for each of the six composite scores in each of the two forms were very high and similar, ranging from 0.89 to 0.96.

Experimental Procedure

After reading the informed consent and confirming their participation in the study, an ad hoc javascript code embedded in the informed consent page randomly assigned students to complete one of the two online surveys corresponding to the two forms of the instrument. Half of the students completed the RP form and the other half the CSEPP form. Presentation of the self-efficacy scales (qualitative and quantitative) within each form was balanced, with half of the students seeing the qualitative scale first and the other half seeing the quantitative scale first. After completing the research self-efficacy items, students were asked to respond to 24 background questions, debriefed, and allowed to submit their personal information for the drawing.

Three factors can potentially affect retrospective pretest scores: item presentation order, item proximity, and item sequencing (Lam, Valencia & Ardeshri, 2014). This study presented the Current ability items before either the Past or Counterfactual items, which is the standard format in retrospective pretest research (Bray et al., 1984; Howard et al., 1979). Item proximity was not-visible, meaning that when responding to a Past or Counterfactual ability item, students could not see their response to the same item regarding their Current ability. The item sequence was across items, meaning that students responded to all Current ability items and then proceeded to respond to all the Past or Counterfactual ability items. Additionally, students could see only one item on the screen at a time, could not revise or change previous responses, and were forced to respond to every item. Other studies (Mueller & Gaus, 2015, 2018; Mueller et al., 2014) did not provide information about how items were presented to participants.

Findings

Descriptive statistics for past and current qualitative, quantitative and overall research self-efficacy scores for the RP and CSEPP conditions are summarized in Table 1.

Table 1
Means (*M*) and Standard Deviations (*SD*) by Treatment Condition and Time

		Qualitative Ability		Quantitative Ability		Overall Ability		
Condition	Time	*M*	*SD*	*M*	*SD*	*M*	*SD*	*N*
Retrospective-Pretest	Before-Program	3.75	1.53	2.52	1.32	3.17	1.13	41
	Within-Program	5.17	1.23	3.48	1.43	4.37	1.05	41
	Overall	4.46	1.55	3.00	1.45	3.77	1.24	82
Counterfactual	Before-Program	3.49	1.47	2.67	1.37	3.10	1.18	40
	Within-Program	4.60	1.41	3.44	1.53	4.05	1.15	40
	Overall	4.05	1.53	3.06	1.49	3.58	1.25	80

A 2x2 mixed ANOVA was conducted with Treatment (RP versus Counterfactual) as the between-subjects variable and Time (Before versus Within Program) as the within-subjects variable. In analyzing the overall research self-efficacy scores, we found a significant gain (M = 1.20) from Before Program to Within Program ($F(1, 79) = 118.59$, $p < .001$). No statistically significant difference was observed for either Treatment main effect ($F(1, 79) = 0.68$, $p = 0.41$) or Time-by-Treatment interaction effect ($F(1, 79) = 1.61$, $p = 0.20$). The main and interaction effects for overall scores are depicted in Figure 1. Similar main and interaction effects were obtained for the qualitative and quantitative subscale scores and these effects are depicted in Figure 2 and 3 respectively.

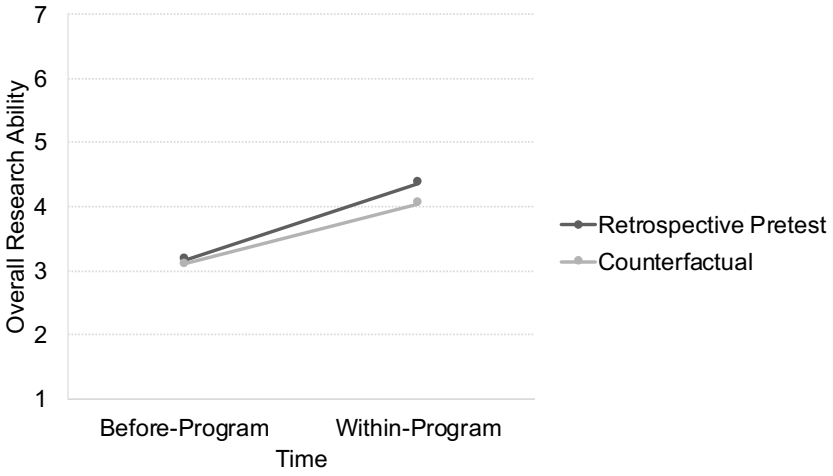

Figure 1. *Means by treatment condition (Retrospective pretest or Counterfactual) and time (before program or within program) for overall research ability scale.*

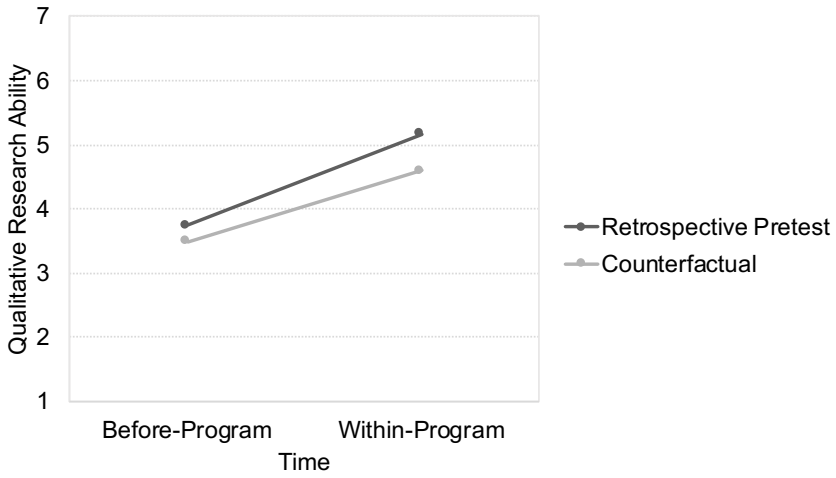

Figure 2. *Means by treatment condition (Retrospective pretest or Counterfactual) and time (before program or within program) for qualitative research ability scale.*

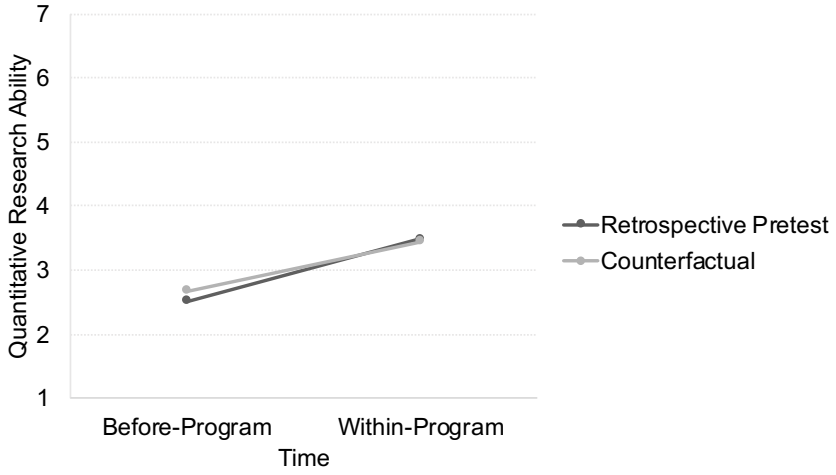

Figure 3. *Means by treatment condition (Retrospective pretest or Counterfactual) and time (before program or within program) for quantitative research ability scale.*

Following the analysis method used by Mueller et al. (2014) and Mueller and Gaus (2015), we computed absolute and relative biases of the treatment effect as well as the 95% confidence interval (CI) using the bootstrap method. According to Mueller and associates, absolute bias (Δ) is the discrepancy between treatment effect or Treatment-on-Treated (TOT) based on retrospective pretest (TOT_{RPRE}) and treatment effect based on CSEPP (TOT_{CSEPP}); relative bias is the proportion of retrospective-pretest-based treatment effect (TOT_{RPRE}) that is due to absolute bias. Treatment effect for each method along with absolute and relative biases are displayed in Table 2.

Table 2
Treatment Effects (TE), Absolute (Δ) and Relative Biases for Retrospective Pretest (RPRE) and CSEPP*

	TOT_{RPRE}			TOT_{CSEPP}			Absolute Bias				Relative
Scale	TE	*p*	*d*	TE	*p*	*d*	Δ	*p*	95% CI		Bias
Qualitative Scale	1.41	0.00	1.02	1.11	0.00	0.77	0.31	0.20	[-0.78	0.16]	28%
Quantitative Scale	0.96	0.00	0.70	0.78	0.00	0.53	0.19	0.44	[-0.66	0.29]	24%
Overall	1.20	0.00	1.10	0.95	0.00	0.82	0.25	0.20	[-0.63	0.13]	26%

*p = p-value of statistical significance tests; d = Cohen's Effect Size coefficient, with 0.2 = small effect, 0.5 = medium effect, and 0.8 large effect; TOT_{RPRE} = difference between "within-program" at the end of the term and "before-program" by the time the student joined the institution; TOT_{CSEPP} = difference between "within-program" at the end of the term and the ability "had the student never joined the institution."

For the *overall* scale scores, absolute bias was 0.25 with 26% relative bias. Absolute and relative biases for the qualitative scale were 0.31 and 28% and for the quantitative scale 0.19 and 24%. All absolute biases were non-significant because the value of zero is included in the respective 95% confidence interval. Relative biases were close to the |25%| cut score rule (Mueller et al, 2014) and deemed tolerable. In comparison, the relative bias for responses to the full scale was 43% in Muller et al.'s 2014 study, which is 21% higher (43% - 22%) than the full scale relative bias from our study. No results were reported for overall or sub-scales in Mueller & Gaus' 2015 study. Mueller et al. (2014) conducted bias analysis at the item level. Because of low reliability of responses to a single item and inflated family-wise error rates from multiple analyses across items (Carifio & Perla, 2007), these findings are not reported.

In sum, differences in treatment effect based on treated participants' report of their counterfactual and the report of pre-program states are practically indistinguishable. The difference is much smaller than the discrepancy in treatment effect comparing treatment participants' counterfactual and a control group's self-report in Mueller and associates' studies.

Discussion

Mueller and Gaus (2015, 2018) and Mueller et al. (2014) advocate and provide evidence in partial support for the use of CSEPP as a replacement of control group self-report data. A significant omission in their research is the direct comparison of CSEPP and retrospective pretest. To fill this void, we conducted a Randomized Control Group experiment in which we compared CSEPP based treatment effect (TOT_{CSEPP}) with retrospective-pretest based treatment effect (TOT_{RPRE}). We found no statistically significant difference between these two forms of treatment effect estimates. Our sample size of 40 and 41 per condition is similar to the average sample size of 36 per condition in Mueller et al.'s study (2014), which also found no statistically significant difference in means between treatment conditions. Although our sample size is not large, results from our analysis appear to suggest that the observed lack of statistical significance is due to the miniscule difference in means between the two treatment conditions.

The proportion of variance of the research efficacy scores explained by treatment conditions (η^2) were 0.008, 0.013, and 0.000 for qualitative, quantitative and overall skills respectively, which, according to Cohen's rule of thumb (1988; Ellis, 2010) are small or trivial. In fact, values of mean scores obtained from the two treatment conditions were practically identical. Between-condition differences in means were 0.13, 0.09, 0.02 for before-program, within-program and overall respectively. The magnitude of these differences was 4%, 2%, and 0.5% of the size of the retrospective pretest means and illustrates the convergence between CSSEP and retrospective pretest responses.

Given the minimal variability in means and percent of variance attributable to differences in treatment conditions, the

observed no-difference in research efficacy between the retrospective-pretest and the counterfactual conditions is unlikely a result of Type 2 error. Additionally, the difference in means was of no *practical* significance; any statistical significance -if detected- would be likely an artifact of a large sample size. That said, an explanation for similarity in findings based on a control group ($T0T_{CG}$) and CSEPP ($T0T_{CSEPP}$) as observed by Mueller and associates is necessary to further our argument that counterfactual estimates and retrospective pretest scores are the same.

Why No Observed Difference in Treatment Effects between CSEPP and Control Group?

While Randomized Control Group experiments can eliminate all threats to internal validity, the Pre-Post-Test-Single-Group and Retrospective-Pre-Post-Test-Single-Group designs are potentially affected by threats to internal validity. If CSEPP is a valid counterfactual, treatment effects estimated by the Randomized Control Group design and the single group CSEPP design should not differ, as it was observed in Mueller and associates' research. On the other hand, if CSEPP is retrospective pretest self-report, the CSEPP design becomes the Retrospective-Pre-Post-Test design; consequently, findings from the Randomized Control Group design and the CSEPP design should differ, *unless the potential internal validity threats in the CSEPP design were not operative* in the particular research setting. In this case, findings from the Randomized Control Group experiment are not expected to differ from that from the CSEPP single group experiment. To test this hypothesis, we systematically reviewed the plausibility of each of the nine internal validity threats being operative in Mueller and associates' control group and CSEPP designs. Prior to reporting findings from our analysis, we would like to set the stage and highlight some key features of their experiments.

The intervention used by Mueller et al. (2014) was about 10 minutes of surfing a webpage that provided "knowledge and recommendations for more climate-friendly consumer behaviour" (p. 12). Mueller and Gaus (2015) asked participants to watch a 30-minute TV documentary that "mainly provided critical and revealing information about the organic food sector in Germany, including organic food production, marketing, and distribution" (p. 9). Directly before and after the webpage surfing or documentary viewing, participants in both the experimental and control groups were asked to complete questionnaires about behavioral intention (Mueller et al., 2014) and about behavior, behavioral intention, and attitude (Mueller & Gaus, 2015). In the 2015 study, the same questionnaire was also administered at a 2-week follow-up.

To examine the likelihood of internal validity threats in Mueller and associates' CSEPP-and-Post-Test design, we employed the procedure proposed by Eckert (2000), which encompasses the situational analysis of evaluation design to determine the plausibility of each threat and consequently ascertain the validity of the treatment effect estimates. Our evaluation of potential internal validity threats in both the control group and CSEPP designs reveals only a potential *testing* effect. Climate control and organic food are salient issues prone to socially desirable responding. Consequently, pretesting might sensitize participants to initiate forming their opinion prior to treatment in both designs (which might be argued as a threat to treatment construct validity). Other threats to internal validity do not appear to be operative in the study by Mueller et al. (2014) and the one by Mueller and Gaus (2015).

We rule out *ambiguous temporal direction* as a threat because treatment precedes the collection of outcome data. The short duration of treatment (10 and 30 minutes) effectively rules out *history* and *maturation* as potential threats. It appears that all participants who completed pretests also completed posttests and CSEPPs, which eliminates the *mortality threat.* Although *instrumentation* is not a threat because the same questionnaire was administered before and after treatment, the aforementioned RSB could be a potential threat in the Pre-Post-Test-Control-group design but not in the CSEPP-Single-Group design, as pre-intervention (CSEPP) scores were collected retrospectively. However, we believe that the general public in Germany would have some familiarity with the treatment content (climate-friendly consumer

behavior and organic food). Therefore, instrumentation related bias as caused by RSB was not likely in the Pre-Post-Test-Control-group design. *Regression* is not a threat because participants in both designs were not recruited based on extreme scores. In regard to *selection* bias, Mueller et al.'s 2014 study used a non-random procedure and Mueller and Gaus (2015) used a randomized procedure to assign participants to control and experimental conditions. In both studies, the authors provided background data to show comparability between groups and conducted covariate analysis to statistically adjust for initial group differences. Selection bias does not appear to be a consequential factor in these between-group designs. Since selection bias does not appear to be a notable validity threat, *interaction* between selection and other internal validity threats is not a factor in the control group designs. In the CSEPP design, selection is not relevant because participants serve in both the experimental and control conditions.

Our analysis reveals only one internal validity threat that was potentially operative in Mueller and associates' two studies published in 2014 and 2015, which could account for the similarity between treatment effects obtained from the control group design and the CSEPP design as reported by Mueller and associates. This observation adds more evidence to findings from our study which repute the distinction between counterfactual thinking and retrospective thinking when crafting pre-intervention status at posttest time. It appears that both counterfactual and retrospective pretest instructions trigger the same mechanism in participants' cognitive process. Estimating pre-intervention status is an act of recalling prior knowledge, whereas counterfactual self-reporting is an act of conjecturing a state that has no real existence. Our findings could suggest that when asked to report a hypothetical condition of not having received treatment, participants were not able to do so and consequently resorted to reporting their positions just before the start of the treatment. Future research should be conducted using cognitive interview (see below) to verify this hypothesis. Even Mueller et. al. acknowledged this possibility as they noted that "respondents may use retrospective thinking for the direct estimation of their own counterfactual. If people had to estimate what they would be like at a certain point in time without having participated in an intervention, they could think about what they were like before the intervention and consider if they are now different from that state" (2014, p. 11).

Bell and Peck (2016) suggested a number of ways to produce counterfactual without control group, including belief, practical experience, extrapolation of prior conditions into the future, and measurement of "non-treated" cases that occur naturally or "deliberately constructed by the researchers through an imposed mechanism or decision rule" (p. 95). Our findings do not support adding counterfactual self-reporting to Bell and Peck's list, not without research to demonstrate strategies and circumstances that could elicit such cognitive skills.

Despite the warning by Mueller and Gaus that "the CSEPP design is not yet ready for use in evaluation practice at the moment" (Mueller & Gaus, 2015, p. 21), the three publications by Mueller and associates could still give evaluators a false sense of confidence and license them to use the methodology, especially in light of the fact that the single group designs are easy to implement. Findings from our study lead us to conclude that until strong evidence is available in support of the validity of counterfactual reported by treated participants, if a single group design is used, evaluators should continue to collect retrospective pretest data instead of counterfactual, and consider the design as a pre- and-post-test single group design with all the potential threats to internal validity embedded in the design to contend with.

Future Research

Our research suggests that counterfactual and retrospective scores are indistinguishable, and counterfactual provided by treatment participants should not be used as a substitute for counterfactual obtained from control group. However, it is only one piece of validity evidence that contradicts the conclusion drawn by Mueller's research. We did not find discriminant validity whereas Mueller found convergent validity of counterfactual provided by the treated. Consequently, more research is needed to determine if counterfactual estimates are just

self-report pretest scores collected retrospectively from treatment participants (retrospective pretest), *or* self-report of the state they would have been in without treatment after they have received the treatment (counterfactual). To settle this debate, it seems logical to tap into respondents' thinking processes. The standard methodology for unravelling cognitive activities is cognitive interview (Willis, 2004; Lam & Valve, In Press). An effective design is to ask participants to report both the pretest and without- treatment states, compare their responses and the cognitive processes underlying these two types of responses, and observe the difficulty participants might experience in differentiating the intent of the two tasks. Other studies could focus on comparing mental activities and validity of self-reports between when participants are asked to contrast two treatments and when they are asked to estimate treatment impact. In addition, Mueller & Gaus (2015) found responses to be more precise with behavioral intention and attitude items than with self-reported behavior. In fact, they later published an article delving into conditions that affect CSEPP (Mueller & Gaus, 2018). These moderating variable effects studies could be replicated with retrospective pretesting to determine if similar findings are observed.

Counterfactual by treatment participants is proposed as an attempt to use self-report to measure treatment impact. In addition to investigating validity of CSSEP estimates, two areas of research pertaining to this proposition seem necessary: the use of self-report to measure treatment impact with retrospective pretesting methodology and the use of the theory-based approach to improve validity of findings from the different retrospective pretest methods.

Self-Report Impact Assessment

Findings regarding validity of self-assessments in general have not been encouraging. For example, Bowman (2010) found that correlations between 1st-year college students' self-report of learning and actual longitudinal gains were small or virtually zero; Nath (2007) found that literacy rates generated through a literacy test were significantly lower than those based on self-report of literacy; and physicians have been shown to be notoriously poor self-assessors (Blanch-Hartigan, 2011; Davis et al., 2006). The common conclusion about inaccuracy of self-assessment is generalizable to situations when self-assessment is used to assess treatment impact.

Using self-report to measure treatment *impact* has been proposed in the business world in the context of Return-on-Investment (ROI) analysis (Phillip, 1996; Phillip & Phillip, 2006; Phillips & Stone, 2002). However, there are serious conceptual flaws, and validity evidence of these measurements is unavailable (Lam, 2008). Also, we noted a study of the validity of self-report of treatment impact showed program staff systematically overestimated program impact on high school students who participated in drug education and prevention programs (Gilham, Lucas, & Sivewright, 1997).

Treatment impact assessment entails estimation of outcomes in the *absence* of treatment through control groups. However, Bell and Peck (2016) point out that "enforcing the embargo on program participation among control group members can be especially challenging and often falls short of universal compliance" (p. 95). One of the challenges is availability of alternative services to control participants. To combat this problem, Bell and Peck (2016) suggest that "randomized experiments with *full* access to alternative services among control group members seems a better approximation to the desired evaluation counterfactual than experiments with *no* control group access to those services" (p. 101). Extrapolating from this observation to self-reporting, instead of asking participants to estimate either a no-treatment state or degree of treatment impact on them, it might be easier and more practical for the participants to compare the target treatment to an alternative treatment, resulting in more valid self-reports.

In the evaluation literature, three retrospective pretesting methods for measuring change have been proposed. These methods are Post plus Retrospective Pretest (P + RP), Post plus Perceived Change (P + PC), and Perceived Change (PC) (Lam & Bengo,

2003). [3] These three self-report methods of change are technically designed to measure *dependent or outcome* variables and not treatment *impact* (Lam, 2009), [4] but, theoretically and with caution, they can be extended to accomplish that. As proposed by Mueller and associates, the P + RP method can be converted to the Post + Counterfactual (P + C) method by rewording a question like "How much do you know just before the intervention?" to "How much would you know without the treatment?" In both the P + PC and the PC methods, a question like "How much have you changed in your knowledge?" could be turned into "How much have you learned from the training?" subsequently creating the Post + Perceived Impact (P + PI) and Perceived Impact (PI) methods. If evaluators are serious about measuring treatment impact with self-report, all the aforementioned retrospective pretesting methods should be examined before implementation.

Theory-Argument-Based Evaluation

Since the use of control groups is rare and self-report is standard in program evaluation, especially in training evaluation, researchers should devote more effort to augment the capacity of the Pre-Post-Test-Single-Group design to assess treatment effect and, generally, to "improve often weak evaluation practice when dealing with causality" (Mayne, 2012, p. 271). Self-reporting of counterfactual by treatment participants is, in our opinion, heading in the wrong direction as it potentially leads us back to the era of input-output black box evaluation (Solmeyer & Constance, 2015). Critical thinking, logical reasoning, and knowledge of content as advocated by theory-driven evaluation (Chen, 1990; Coryn & Hobson, 2011; Ford & Weissbein, 1997; Lipsey, 1993; Rogers & Weiss, 2007; White, 2009) and validity-argument-based evaluation (Cronbach & Meehl, 1955; Kane, 2006, 2013; LeBaron Wallace, 2011; Messick, 1989; Peck, Kim, & Lucio, 2012) should lead the way instead, albeit theory-argument-based evaluation and experimentation are not mutually exclusive (Cook, 2002).

Over three decades ago, the message arising from the monograph edited by Trochim (1986) called for a move from a mechanistic *cookbook* quasi-experimental or method-driven evaluation (Chen, 1990) to a more integrated and synthetic approach that is both theoretical and contextualized. This advice still resonates today. Additionally, the single-group experimental designs are plagued with internal validity threats as discussed earlier in this paper, "the more threats to internal validity there are, the more important treatment theory becomes" (Lipsey, 1993, p. 49). "Theory-based evaluation could strengthen the validity of evaluations when random assignment is impossible" (Weiss, 1997, p. 43).

The theory or argument based approach to assessing treatment impact embraces treatment impact assessment strategies that already exist in the evaluation literature, including contribution analysis (Mayne, 2001, 2012), causal mediation analysis (Keele, 2015), rival hypothesis methods (Rindskopf, 2000; Yin, 2000), ruling out validity threats [as shown in the previous section] (Eckert, 2000; Reichardt, 2000), process tracing (Schmitt & Beach, 2015), pattern matching (Shadish et al, 2002), realistic evaluation (Blamey & Mackenzie, 2007; Pawson & Tilley, 2001; Porter & O'Halloran, 2011), and correspondence between treatment utilization and outcomes (e.g., Solmeyer & Constance, 2015).

Back in 1993, Lipsey found that fewer than 10% of the research articles he reviewed presented some theoretical context. Though we don't know what the current trend is, suffice it to say that future development in impact evaluation should aim to reverse this trend and perhaps fulfil the vision Lipsey laid out for us that "every report of a treatment effectiveness study includes a section labelled 'treatment theory'" (Lispsey & Wison, 1993, p. 58).

In closing, self-report is the most efficient and consequently the most popular data in program evaluation, especially short-term interventions like training workshops.

[3] Since the PC method does not require participants to report pretest status, *Retrospective Self-report of Change* (RSRC) might be more appropriate than *Retrospective Pretest* to label this form of self-reporting.

[4] Practitioners often misused self-report data to measure treatment effect, partially due to the confusion between outcomes and impacts (Belcher & Palenberg, 2018) and the low accuracy of self-report performance data.

However, its usage in program evaluation, including the aforementioned retrospective pretest methods, is not commensurate with the attention paid to its validity (Lam & Bengo, 2003; Hill & Betz, 2005; Nimon, Zigarmi, & Allen, 2011; Pratt, McGuigan, & Katzev, 2000; Taylor, Russ-Eft, & Taylor, 2009). We hope that our research sparks greater interest in assessing validity of findings based on retrospective pretesting methods, as well as in utilizing theory or validity-argument to evaluate treatment impact.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Belcher, B., & Palenberg, M. (2018). Outcomes and Impacts of Development Interventions: Toward Conceptual Clarity. *American Journal of Evaluation*, 1098214018765698. https://doi.org/10.1177/1098214018765698

Bell, S.H. & Peck, L.R. (2016). On the feasibility of extending social experiments to wider applications. *Journal of MultiDisciplinary Evaluation, 12*(27), 2016

Blamey, A., & Mackenzie, M. (2007). Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges? *Evaluation*, *13*(4), 439–455. https://doi.org/10.1177/1356389007082129

Blanch-Hartigan, D. (2011). Medical students' self-assessment of performance: results from three meta-analyses. *Patient Education and Counseling*, *84*(1), 3–9. https://doi.org/10.1016/j.pec.2010.06.037

Bowman, N. (2010). Can 1st-Year College Students Accurately Report Their Learning and Development? *American Educational Research Journal*, *47*(2), 466–496. https://doi.org/10.3102/0002831209353595

Bray, J. H., Maxwell, S. E., & Howard, G. S. (1984). Methods of Analysis with Response-Shift Bias. *Educational and Psychological Measurement*, *44*(4), 781–804.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Campbell, D. T., & Stanley, J. (1963). *Experimental and Quasi-Experimental Designs for Research* (1 edition). Boston: Wadsworth Publishing.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3*(3), 106–116.

Chen, H.-T. (1990). *Theory-driven evaluations.* Newbury Park: Sage.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 edition). Hillsdale, N.J: Routledge.

Cook, T. D. (2002). Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community Has Offered for Not Doing Them. *Educational Evaluation and Policy Analysis*, *24*(3), 175–199.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design & analysis issues for field settings*. Rand McNally College Pub. Co.

Coryn, C. L. S., & Hobson, K. A. (2011). Using nonequivalent dependent variables to reduce internal validity threats in quasi-experiments: Rationale, history, and examples from practice. *New Directions for Evaluation*, *2011*(131), 31–39. https://doi.org/10.1002/ev.375

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.https://doi.org/10.1037/h0040957

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA: The Journal of the American*

Medical Association, *296*(9), 1094–1102. https://doi.org/10.1001/jama.296.9.1094

Dennis, M. L. (1990). Assessing the Validity of Randomized Field Experiments: An Example from Drug Abuse Treatment Research. *Evaluation Review*, *14*(4), 347–373. https://doi.org/10.1177/0193841X9001400402

Eckert, A. (2000). Situational enhancement of design validity: the case of training evaluation at the World Bank Institute. *The American Journal of Evaluation*, *21*(2), 185–193. https://doi.org/10.1016/S1098-2140(00)00065-5

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press

Ford, J. K., & Weissbein, D. A. (1997). Transfer of Training: An Updated Review and Analysis. *Performance Improvement Quarterly*, *10*(2), 22–41. https://doi.org/10.1111/j.1937-8327.1997.tb00047.x

Gilham, S. A., Lucas, W. L., & Sivewright, D. (1997). The Impact of Drug Education and Prevention Programs: Disparity Between Impressionistic and Empirical Assessments. *Evaluation Review*, *21*(5), 589–613. https://doi.org/10.1177/0193841X9702100504

Hill, L. G., & Betz, D. L. (2005). Revisiting the Retrospective Pretest. *American Journal of Evaluation*, *26*(4), 501–517. https://doi.org/10.1177/1098214005281356

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945. https://doi.org/10.2307/2289064

Howard, G. S. (1980). Response-Shift Bias A Problem in Evaluating Interventions with Pre/Post Self-Reports. *Evaluation Review*, *4*(1), 93–106. https://doi.org/10.1177/0193841X8000400105

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, *64*(2), 144–150.

Howard, G. S., Schmeck, R. R., & Bray, J. H. (1979). Internal Invalidity in Studies Employing Self-Report Instruments: A Suggested Remedy. *Journal of Educational Measurement*, *16*(2), 129–135. https://doi.org/10.1111/j.1745-3984.1979.tb00094.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed, pp. 17–64). Westport, CT: Praeger Publishers.

Keele, L. (2015). Causal Mediation Analysis: Warning! Assumptions Ahead. *American Journal of Evaluation*, *36*(4), 500–513. https://doi.org/10.1177/1098214015594689

Lam, T.C.M. (2008). *Estimating program impact through judgment: A simple but bad idea?* Paper presented at the annual meeting of the American Evaluation Association.

Lam, T.C.M. (2009). Do Self-Assessments Work to Detect Workshop Success? An Analysis of D'Eon et al.'s Argument and Recommendation. *American Journal of Evaluation, 30* (1)93-105.

Lam, T.C.M. & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practices. *American Journal of Evaluation, 24*(1), 65- 80.

Lam, T.C.M., Valencia, E., & Ardeshri, M. (2014). *Item order effect in retrospective pretest method of self-reporting change.* Paper presented at the annual meeting of the American Evaluation Association.

Lam, T.C.M. & Valve, L. (In Press). Cognitive Interview in Survey Development: Item Construction and Response Validation. In Harbaugh, G. & Luhanga, U. (Ed.) *Basic Elements of Survey Research in Education: Addressing the Problems Your Advisor Never Told You About.* American Educational Research Association (AERA).

LeBaron Wallace, T. (2011). An argument-based approach to validity in evaluation. *Evaluation*, *17*(3), 233 246. https://doi.org/10.1177/1356389011410522

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. *New Directions for*

Program Evaluation, *1993*(57), 5–38. https://doi.org/10.1002/ev.1637

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*(12), 1181–1209. https://doi.org/10.1037/0003-066X.48.12.1181

Mayne, J. (2001). Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly. *Canadian Journal of Program Evaluation*, *16*(1), 1–24.

Mayne, J. (2012). Contribution analysis: Coming of age? *Evaluation*, *18*(3), 270–280. https://doi.org/10.1177/1356389012451663

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–110). New York, NY: MacMillan.

Mills, B. A., Caetano, R., & Rhea, A. E. (2014). Factor Structure of the Clinical Research Appraisal Inventory (CRAI). *Evaluation & the Health Professions*, *37*(1), 71–82. https://doi.org/10.1177/0163278713500303

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*(1), 7–24. https://doi.org/10.1177/1098214014538487

Mueller, C. E., & Gaus, H. (2018). Treatment Effect Estimation Using Self-Estimated Counterfactuals Under Varying Conditions. *Journal of MultiDisciplinary Evaluation*, *14*(30), 16–36.

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*(1), 8–25. https://doi.org/10.1177/1098214013503182

Mullikin, E. A., Bakken, L. L., & Betz, N. E. (2007). Assessing Research Self-Efficacy in Physician-Scientists: The Clinical Research APPraisal Inventory. *Journal of Career Assessment*, *15*(3), 367–387. https://doi.org/10.1177/1069072707301232

Nath, S. R. (2007). Self-Reporting and Test Discrepancy: Evidence from a National Literacy Survey in Bangladesh. *International Review of Education / Internationale Zeitschrift Für Erziehungswissenschaft / Revue Internationale de l'Education*, *53*(2), 119–133.

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*(1), 8–28. https://doi.org/10.1177/1098214010378354

Pawson, R., & Tilley, N. (2001). Realistic Evaluation Bloodlines. *American Journal of Evaluation*, *22*(3), 317–324. https://doi.org/10.1177/109821400102200305

Peck, L. R., Kim, Y., & Lucio, J. (2012). An Empirical Examination of Validity in Evaluation. *American Journal of Evaluation*,*33*(3), 350–365. https://doi.org/10.1177/1098214012439929

Phillip, J. J. (1996). Measuring ROI: the fifth level of evaluation. *Technical & Skills Training*, (April), 10–13.

Phillip, J. J., & Phillip, P. (2006). Measuring return of investment in leadership development. In K. Hannum, J. W. Martineau, & C. Reinelt (Eds.), *The Handbook of Leadership Development Evaluation*. John Wiley & Sons.

Phillips, J., & Stone, R. (2002). *How to Measure Training Results : A Practical Guide to Tracking the Six Key Indicators* (1 edition). New York: McGraw-Hill Education.

Porter, S., & O'Halloran, P. (2011). The use and limitation of realistic evaluation as a tool for evidence-based practice: a critical realist perspective. *Nursing Inquiry*, *19*(1), 18–28. https://doi.org/10.1111/j.1440-1800.2011.00551.x

Pratt, C. C., McGuigan, W. M., & Katzev, A. R. (2000). Measuring Program Outcomes: Using Retrospective Pretest Methodology. *American Journal of Evaluation*, *21*(3), 341–349.

https://doi.org/10.1177/109821400002100305

Reichardt, C. S. (2000). A Typology of Strategies for Ruling Out Threats to Validity. In L. Bickman (Ed.), *Research Design: Donald Campbell's Legacy* (1 edition). Thousand Oaks, Calif: SAGE Publications, Inc.

Rindskopf, D. (2000). Plausible rival hypotheses in measurement, design and scientific theory. In L. Bickman (Ed.), *Validity and social experimentation*. Sage Publications.

Rogers, P. J., & Weiss, C. H. (2007). Theory-based evaluation: Reflections ten years on: Theory-based evaluation: Past, present, and future. *New Directions for Evaluation*, (114), 63–81. https://doi.org/10.1002/ev.225

Schmitt, J., & Beach, D. (2015). The contribution of process tracing to theory-based evaluations of complex aid instruments. *Evaluation*, *21*(4), 429–447. https://doi.org/10.1177/1356389015607739

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.

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

Solmeyer, A. R., & Constance, N. (2015). Unpacking the "Black Box" of Social Programs and Policies: Introduction. *American Journal of Evaluation*, *36*(4), 470–474. https://doi.org/10.1177/1098214015600786

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Gilding the Outcome by Tarnishing the Past Inflationary Biases in Retrospective Pretests. *American Journal of Evaluation*, *30*(1), 31–43. https://doi.org/10.1177/1098214008328517

Trochim, W. M. K. (1986). Editor's notes. *New Directions for Program Evaluation*, (31), 1–7. https://doi.org/10.1002/ev.1430

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation*, (76), 41–55. https://doi.org/10.1002/ev.1086

White, H. (2009). Theory-based impact evaluation: principles and practice. *Journal of Development Effectiveness*, *1*(3), 271–284. https://doi.org/10.1080/19439340903114628

Willis, G. B. (2004). *Cognitive Interviewing: A Tool for Improving Questionnaire Design* (1st ed.). Sage Publications, Inc.

Yin, R. K. (2000). Rival Explanations as an Alternative to Reforms as "Experiments." In L. Bickman (Ed.), *Validity & social experimentation* (Vol. 1). Thousand Oaks: SAGE Publications.