Published in final edited form as:

Res Nurs Health. 2015 June ; 38(3): 241–247. doi:10.1002/nur.21651.

Assessment Effects in Educational and Psychosocial Intervention Trials: An Important but Often-Overlooked Problem

Mi-Kyung Song and

Associate Professor, Beerstecher and Blackwell Term Distinguished Scholar, School of Nursing, University of North Carolina at Chapel Hill, 7460 Carrington Hall, Chapel Hill, NC 27599

Sandra E. Ward

Helen Denne Schulte Professor Emerita, School of Nursing, University of Wisconsin-Madison, Madison, WI

Mi-Kyung Song: songm@email.unc.edu

Abstract

Baseline assessments and repeated measures are an essential part of educational and psychosocial intervention trials, but merely measuring an outcome of interest can modify that outcome, either by the measurement process alone or by interacting with the intervention to strengthen or weaken the intervention effects. Assessment effects can result in biased estimates of intervention effects and may not be controlled by the usual two-group randomized controlled trial design. In this paper, we review the concept of assessment effects and other related phenomena, briefly describe study designs that estimate assessment effects separately from intervention effects and discuss their strengths and limitations, review evidence regarding the strength of assessment effects in intervention trials targeting behavior change, and discuss implications for intervention research.

Keywords

Assessment effects; measurement; behavior change; randomized trials; interventions; bias; internal validity; external validity

In 1908, Winch conducted the first two-group experimental study, in which the experimental group received pre-test, intervention, and post-test, while the control group received the pre-test, no intervention, and the post-test. He argued that without a control group (that is, with a single group pre-experimental design), it is not possible to determine whether changes in dependent variables are due to subjects taking the pre-test or to the intervention itself. Ever since, this two-group experimental design has been the gold standard for testing interventions.

However, it is not uncommon that seemingly well-thought-out two-group randomized controlled trials (RCTs) of educational and psychosocial interventions fail to demonstrate a

No authors have conflict of interest or financial relationships to disclose.

significant intervention effect on targeted outcomes when compared to the control condition. Such a finding can simply suggest that the intervention is not, in fact, any better than the control condition, but the null finding also could be due to insufficient intervention strength (a potentially good intervention, but just not enough of it), lack of intervention fidelity or adherence to the treatment regimen, outcome measures that are not sensitive, inadequate timing of outcome assessment, or lack of statistical power. Furthermore, we often observe that not only do participants in the intervention group show improvement in outcomes, but so do those in the control group (McCambridge & Kypri, 2011). Due to failure to detect a significant therapeutic effect above that seen in the control group, many well-designed interventions that could potentially alter the course of illness or quality of life of patients and their family members do not move forward.

Assessments can act as an intervention, resulting in the control group improving roughly as much as does the intervention group (McCambridge, Butor-Bhavsar, Witton, & Elbourne, 2011; McCambridge & Kypri, 2011; McCambridge, Kypri, & Elbourne, 2014), or may interact with an intervention to inflate the intervention effects, resulting in erroneous conclusions that the intervention is superior to control. This phenomenon is addressed in design textbooks but often overlooked in actual trials and has received little attention in nursing. The purposes of this paper are to review the concept of assessment effects with attention to related phenomena, to briefly describe study designs that estimate assessment effects and discuss their strength and limitations, to review evidence regarding the strength of assessment effects in intervention trials, and to discuss implications for intervention research.

Assessment Effects and Related Phenomena

In the following paragraphs, we discuss assessment effects and two very closely related phenomena, reactivity and pre-test sensitization. We then discuss two phenomena, self-monitoring and the Hawthorne effect, that are not quite as closely related but that show substantial overlap. (Depending on the discipline, these phenomena may be referred to using different terminology.) We do not address even more distantly-related phenomena that are important in trial design, such as placebo effects, demand characteristics, and non-specific effects, as these issues are beyond the scope of this paper.

Assessment effects, first described by Solomon (1949) and further discussed by Campbell (1957), refer to a phenomenon in which the outcome of interest (e.g., a behavior, an attitude, or knowledge) is modified by merely assessing it. The effect can be exerted either by the assessment alone or by its interaction with the intervention to either strengthen or weaken the intervention effects. The exact mechanism by which assessments have an effect is not fully understood but is thought to occur by raising subjects' awareness (in which case, assessments may affect the outcomes for both intervention and control groups) or by changing the subjects' perception of the intervention (Solomon, 1949).

Campbell and Stanley (1963) and Cook and Campbell (1979) later separated assessment effects into testing effects and reactivity. In their definition, testing effects refer to the effects of taking a test on the outcomes of taking a second test, and reactivity refers to a

phenomenon in which a pre-test either increases or decreases a subject's sensitivity or responsiveness to the intervention. However, the distinction between testing effects and reactivity is blurred in the literature. For example, Shadish, Cook and Campbell (2002) described testing effects as "a form of reactivity" (p.60) that may occur "when research participants are provided with cues to features of treatment" (p. 78). Others have used the term assessment reactivity to refer to

> …the finding that the action of having a behavior queried, monitored, or become a focus of attention during a research study independently affects the expression of that behavior regardless of other interventions or manipulations used in the study (Schrimsher & Filtz, 2011, p. 108).

Assessment effects can threaten internal validity in a wide variety of studies, not only those in which self-report is the sole method of assessing outcomes. Even objective assessment can cause reactivity in a participant. For example, measuring a participant's height is relatively nonreactive (unlikely to cause reactivity in the participant), but measuring weight as a baseline assessment in a study of a weight control intervention may well be reactive and could stimulate weight reduction with or without the intervention (Campbell, 1957).

Pre-test sensitization, another term for assessment effects, is defined as "the potential or actuality of a pretreatment assessment's effect on subjects in an experiment" (Willson & Kim, 2010, p.1092). Researchers since the early 1960s in the fields of cognitive psychology, developmental psychology, and education have demonstrated the presence of such effects (Lipowski, Pyc, Dunlosky, & Rawson, 2014; Roediger & Karpicke, 2006; van den Broek, Segers, Takashima, & Verhoeven, 2013). In education, one may give a pre-test of knowledge on a certain topic before lecturing on that topic. The pre-test shows students that they have a great deal to learn, helps focus their attention, and sensitizes them to content that is provided in the lecture. RCTs of educational interventions in health-related trials commonly achieve significant improvements knowledge or attitudes. The significant improvement in those outcomes post-intervention could be a function of the combination of pre-test sensitization and intervention effects, rather than of intervention effects alone.

In self-monitoring, a related concept in psychology and educational psychology, a person is asked to observe and record some aspect of his/her thoughts, feelings, or behavior (Cole & Bambara, 2000). "Simply the act of engaging in self-monitoring without any additional intervention components can lead to changes in the targeted behavior" (Cole & Bambara, p. 203). The mechanism by which self-monitoring changes behavior may be the self-awareness that arises from self-observation and self-recording, which becomes feedback on one's performance (Cole & Bambara). Repeated self-assessments over time are an effective behavior change technique (French et al., 2008), as have been demonstrated in school settings since the early 1970s, in children who have learning disabilities or attention deficit or hyperactivity disorders (Cole & Bambara). Self-monitoring also has reduced problem behaviors, such as driving while intoxicated (Neff & Landrum, 1983; Sanchez-Craig, Davila, & Cooper, 1996) and sexual risk-taking (Weinhardt, Carey, & Carey, 2000). Repeated assessments over time are often a necessary component of efficacy trials, not as an intended intervention but rather to measure the intervention effect over time. In those

situations, repeated assessments threaten valid interpretations about the efficacy of the intervention being tested.

In the field of neuropsychology, the term practice effects is used to refer to a phenomenon in which participants learn a test-taking strategy from repeated exposure to the testing, such as neuropsychological assessments (Basso, Carona, Lowery, & Axelrod, 2002; Beglinger et al., 2005; Benedict & Zgaljardic, 1998). Alternate forms of the tests are commonly used to reduce practice effects ("noise") in detecting "true" cognitive functioning and have shown to be effective in tests of some cognitive domains (Crawford, Stewart, & Moore, 1989; Zgaljardic & Benedict, 2001). The alternate form must assess the same construct equally well. While alternate forms may reduce practice effects when compared to use of the same form, such a reduction may be moderated by test subject. For example, novel tests (i.e., tasks unlikely to have been performed in everyday life by participants) and those with a large cognitive demand have the greatest practice effects (Beglinger et al. ; Benedict & Zgaljardic). Furthermore, in most educational and psychosocial interventions, the act of answering questions regardless of test form is likely to cause reactivity.

Finally, a concept somewhat different from the assessment effects discussed above is the Hawthorne effect, so named because it was first discovered in a series of experiments conducted at an industrial plant in Hawthorne, Illinois. The studies, each of which was a basic single-group pre-post design, were intended to determine whether alterations of the work environment increased worker productivity. It was discovered that *any* change seemed to have the intended effect. The changes in productivity seemed to have nothing to do with physical working conditions but rather to the attention or concern being paid to the workers. Hence, the Hawthorne effect has come to mean effects on outcomes that are due to subjects' awareness of being studied (Mayo, 1933; Parsons, 1974). Since those early studies, further work on the Hawthorne effect has led to wide disagreements on its mechanisms of action and on what the Hawthorne effect actually is (McCambridge, Witton, & Elbourne, 2014). For example, Shadish, Cook and Campbell (2002) described the Hawthorne effect as one example of "novelty and disruption effects" (p.79) that occur when observation is introduced. They were critical of the early interpretation that the Hawthorne effect is due to attention being given to the subjects and saw the effect as due to novelty. Some investigators in medicine have returned to the idea that the Hawthorne effect refers to subjects' awareness of being observed but results from the combined effects of observation and assessment (Kaptchuk et al., 2008). Working from this perspective, Kaptchuk and colleagues argued that clinicians can make use of these effects to amplify the therapeutic effects of an intervention and maximize clinical outcomes.

In the paragraphs above, we have alluded to several potential mechanisms by which assessments may have an impact on outcomes. That pre-tests in a classroom help focus students' attention on specific material seems to be a face-valid and logical explanation, but we are not aware of studies to define, operationalize, measure, or test the mechanisms by which assessment effects operate.

Designs to Quantify Assessment Effects

Pre-test sensitization, self-monitoring, and more recently the Hawthorne effect have been seen as useful in the clinical application of therapeutic interventions. Nonetheless, controlling or testing assessment effects is important when intervention efficacy is being evaluated. Without valid estimates of true intervention effects, it is difficult to replicate research and to translate findings into clinical settings. Once an intervention is found to be efficacious, one might then wish to increase its potency by using these related phenomena.

To control for assessment effects during intervention trials, one needs to return to Solomon's work. In a first effort to quantify assessment effects, Solomon (1949) expanded the two-group experimental design to include a third group (See Table 1). This design includes an intervention group and a no-intervention control group, both of which complete pre-tests, and a third group that receives no pre-test but receives the intervention. All groups complete post-tests. With the additional control group, the pre-test effect alone and the intervention effect alone can be estimated. However, the effects of the interaction between assessment and intervention cannot be estimated with the three-group design.

Solomon then expanded the design further to include four groups: 1) an intervention group receiving the pre-test and intervention, 2) a control group receiving a pre-test but no intervention, 3) a control group receiving no pre-test and receiving the intervention, and 4) a control group receiving neither pre-test nor intervention (Table 1). This simple 2 × 2 factorial design enables the investigator to assess the main effects of assessment and the main effects of intervention as well as the interaction between assessment and intervention. This design allows a clear comparison of the effects of the intervention with and without the presence of assessment effects. However, as Solomon (1949) noted, the estimated assessment effects alone may not be the true assessment effects, especially in trials conducted over several years, because other factors, such as maturation or growth, could be affecting post-test scores.

Despite their utility to quantify and thus control for assessment effects, Solomon designs have not been widely used. Their lack of popularity in the actual conduct of research may be due to pragmatic issues, such as the larger sample size, time, and costs of such a trial compared to a two-group trial. Furthermore, while Solomon's three- or four-group designs control pre-test effects, these designs do not control for the effects of repeated assessments of outcome variables. In addition, other factors that cause reactivity, such as research activities that occur prior to randomization (e.g., screening or consenting processes) and randomization itself (Brewin & Bradley, 1989; King et al., 2005; Zelen, 1990) cannot be controlled with three- and four-group designs.

The need for designs that separate assessment effects from intervention effects has been recognized in addiction research, in which screening and comprehensive assessment of drug and alcohol use are required for treatment. Investigators have recognized that such extensive assessments may mask the potency of the intervention under study (Daeppen et al., 2007; Donovan et al., 2012; Epstein et al., 2005; Schrimsher & Filtz, 2011). Recently, Donovan and colleagues (2012) have proposed a three-group design to examine assessment effects on

drug and alcohol abuse behaviors, the Screening, Motivational Assessment, Referral, and Treatment in Emergency Departments protocol. In this design, all potential subjects are screened, and individuals identified as positive cases are invited to join the study. Subjects are then randomly assigned to one of three conditions: minimal screening only, screening + baseline assessment, or screening + baseline assessment + intervention. Subjects in all three groups complete repeated measures at 3, 6, and 12 months. This design allows investigators to estimate baseline assessment effects (by comparing screening alone to screening + baseline assessment). One can also determine the effects of baseline assessment + intervention (by comparing screening + baseline assessment + intervention to screening + baseline assessment). However, because this design does not include a group receiving screening + intervention, the intervention effect without the effect of baseline assessment cannot be estimated, and it cannot determine the screening effects alone, because all groups receive the initial screening. Further, the effects of repeated measures cannot be examined in this design. Nonetheless, this innovative design may prove to be useful in studies wherein potentially dangerous behaviors require rapid screening followed by a comprehensive baseline assessment.

Despite the problems that assessment can cause, one must remember that any effort to design a study to control or at least measure the impact of assessment effects must take into account the essential role these assessments play. Baseline assessment is intended to demonstrate the counterfactual (what happens in a group not exposed to the intervention). A design such as the Solomon 4-group accommodates this need, and any other new designs intended to measure or control assessment effects must similarly address the need to demonstrate the counterfactual.

Assessment Effects in Intervention Trials Targeting Behavior Change

A large body of evidence of the effects of assessment has arisen from its use as an intervention strategy to improve outcomes, such as memory, learning, or risk behaviors. In those studies, the phenomena, such as self-monitoring, was the actual intended intervention rather than an extraneous factor that could influence outcomes or interact with the intervention being tested. To examine whether simply asking a few questions about a target behavior (as a control condition) could change the behavior, McCambridge and Kypri (2011) conducted a systematic review of ten traditional two-group RCTs of brief alcohol interventions, in which control groups completed a few simple questions about drinking behaviors while intervention groups completed a more comprehensive assessment of alcohol consumption. All of the reports indicated improvement in both control and intervention groups, but the standardized effect sizes of assessment varied substantially ($z = 0.14 – 2.46$) depending on the targeted behavior (e.g., weekly vs. daily alcohol consumption), sample (e.g., college students vs. clinic patients), and setting (e.g., emergency department or other health care setting vs. university campus). Assessment effects were stronger in trials targeting weekly alcohol consumption; their effect size was approximately 35% of the known effect of brief alcohol interventions in primary care settings (Kaner et al., 2007). Assessment effects were also stronger in trials targeting students and in trials conducted outside of healthcare settings.

Only a few behavior change intervention studies using the Solomon four-group design to control assessment effects have been reported, probably for the reasons described above (e.g., cost). Nonetheless, McCambridge and colleagues (2011) reviewed ten studies that used the Solomon four-group design for interventions targeting various behaviors (e.g., contraception use, drinking behaviors, and physical activity). As in their previous systematic review, the individual studies unfortunately were too heterogeneous in design, sample, setting, length of follow-up, and outcome measures to allow for meta-synthesis, but four showed evidence of main effects of assessment on self-reported behavioral outcomes (Campanelli, Dielman, Shope, Butchart, & Renner, 1989; van Sluijs, van Poppel, Twisk, & van Mechelen, 2006) and on knowledge and intention outcomes (Dignan et al., 1998; Duryea, 1983). For instance, in a study by van Sluijs and colleagues, the odds of meeting recommended levels of physical activity at 6 months were notably higher in those who completed baseline assessment than in those who did not (*OR* = 1.7; 95% CI 1.14 – 2.54).

The only trial in the McCambridge's review (2011) that demonstrated the presence of an interaction between assessment and intervention was by Kvalem (1996), in which the intervention effect was significantly greater for participants who completed baseline assessment compared to those who received the intervention without baseline assessment. It should be noted that the sample sizes in the other trials in this review were generally quite large (*N* = 717 to 5680). Nonetheless, without careful power analyses, one cannot presume that they had sufficient power to detect interactions. Taken together, these systematic reviews suggest that, although the size of assessment effects in intervention trials vary, uncontrolled assessment effects can undermine an investigator's ability to make accurate inferences about intervention effects.

Implications for Intervention Research

When a trial of an educational or psychosocial intervention fails to demonstrate significant intervention effects, investigators search for an explanation for the null finding, such as looking for any systematic biases that might have been overlooked, and discuss what they might do differently in future research. On the other hand, when the opposite happens -- that is, when an intervention is found to be significantly superior to control -- then the intervention typically receives all of the credit for the positive results, and investigators seldom reflect on what else, other than the intervention, might have contributed to the positive outcomes.

The implications of failing to reflect on positive results can be manifold. Consider that some interventions seem to "work" in the study context but then fail to be effective in practice. While there are many factors affecting implementation (Durlak & DuPre, 2008), one reason for such failure may be that those interventions are no longer accompanied by the assessments that were present in the efficacy trials. Therefore, estimating intervention effects without understanding assessment effects may delay translating research findings into practice, a consequence that can have immense implications for patients' well-being and for the costs of the trials themselves.

As described above, assessment effects either alone or in concert with an intervention may influence the outcome of a trial, and may either deflate or inflate the intervention effect. Therefore, even when there are significant differences in outcomes between the intervention and control groups, this does not rule out the possibility that assessment effects are operating. Similarly, a finding of no difference between the intervention and control groups may not indicate that the intervention has no effect. Unfortunately, while considerable effort has been made to develop designs that control assessment effects in intervention trials, these designs can control some, but not all, assessment effects.

Given the lack of methods to quantify assessment effects and unclear mechanisms of action, researchers should carefully interpret study findings regardless of whether they are positive or null. The literature on assessment effects (and other related phenomena) reminds us that we have only a modest understanding of what is going on with participants in intervention trials; we know little about the research conditions that may cause participants' reactivity and that may impede valid inferences about the intervention effects. It is clear that further investigations into optimal methodologies for intervention research are needed.

Acknowledgments

This work was supported by a grant from the National Institutes of Health (R01NR011464, PI: Song).

References

Basso MR, Carona FD, Lowery N, Axelrod BN. Practice effects on the WAIS-III across 3- and 6-month intervals. Clincal Neuropsychology. 2002; 16:57–63.10.1076/clin.16.1.57.8329

Beglinger LJ, Gaydos B, Tangphao-Daniels O, Duff K, Kareken DA, Crawford J, Siemers ER. Practice effects and the use of alternate forms in serial neuropsychological testing. Archives of Clinical Neuropsychology. 2005; 20:517–529.10.1016/j.acn.2004.12.003 [PubMed: 15896564]

Benedict RH, Zgaljardic DJ. Practice effects during repeated administrations of memory tests with and without alternate forms. Journal of Clinical Experimental Neuropsychology. 1998; 20:339–352.10.1076/jcen.20.3.339.822 [PubMed: 9845161]

Brewin CR, Bradley C. Patient preferences and randomised clinical trials. British Medical Journal. 1989; 299(6694):313–315. [PubMed: 2504416]

Campanelli PC, Dielman TE, Shope JT, Butchart AT, Renner DS. Pretest and treatment effects in an elementary school-based alcohol misuse prevention program. Health Education Quarterly. 1989; 16:113–130. [PubMed: 2703348]

Campbell, D.; Stanley, J. Experimental and quasi-experimental designs for research. Chicago, IL: Rand-McNally; 1963.

Campbell DT. Factors relevant to the validity of experiments in social settings. Psychological Bulletin. 1957; 54:297–312. [PubMed: 13465924]

Cole, CL.; Bambara, LM. Self-monitoring: Theory and practice. In: Kratochwill, TR.; Shapiro, ES., editors. Behavioral assessment in schools: Theory, research and clinical foundations. 2. New York, NY: The Guilford Press; 2000. p. 202-232.

Cook, TD.; Campbell, DT. Quasi-experimentation: Design and analysis for field settings. Boston, MA: Houghton Mifflin Co; 1979.

Crawford JR, Stewart LE, Moore JW. Demonstration of savings on the AVLT and development of a parallel form. Journal of Clinical Experimental Neuropsychology. 1989; 11:975–981.10.1080/01688638908400950 [PubMed: 2592534]

Daeppen JB, Gaume J, Bady P, Yersin B, Calmes JM, Givel JC, Gmel G. Brief alcohol intervention and alcohol assessment do not influence alcohol use in injured patients treated in the emergency

department: A randomized controlled clinical trial. Addiction. 2007; 102:1224–1233.10.1111/j.1360-0443.2007.01869.x [PubMed: 17565563]

Dignan MB, Michielutte R, Wells HB, Sharp P, Blinson K, Case LD, McQuellon RP. Health education to increase screening for cervical cancer among Lumbee Indian women in North Carolina. Health Education Research. 1998; 13:545–556. [PubMed: 10345905]

Donovan DM, Bogenschutz MP, Perl H, Forcehimes A, Adinoff B, Mandler R, Walker R. Study design to examine the potential role of assessment reactivity in the Screening, Motivational Assessment, Referral, and Treatment in Emergency Departments (SMART-ED) protocol. Addiction Science & Clinical Practice. 2012; 7:16.10.1186/1940-0640-7-16 [PubMed: 23186329]

Duryea EJ. Utilizing tenets of inoculation theory to develop and evaluate a preventive alcohol education intervention. Journal of School Health. 1983; 53:250–256. [PubMed: 6552340]

Epstein EE, Drapkin ML, Yusko DA, Cook SM, McCrady BS, Jensen NK. Is alcohol assessment therapeutic? Pretreatment change in drinking among alcohol-dependent women. Journal of Studies on Alcohol. 2005; 66:369–378. [PubMed: 16047526]

French DP, Wade AN, Yudkin P, Neil HA, Kinmonth AL, Farmer AJ. Self-monitoring of blood glucose changed non-insulin-treated Type 2 diabetes patients' beliefs about diabetes and self-monitoring in a randomized trial. Diabetic Medicine. 2008; 25:1218–1228.10.1111/j.1464-5491.2008.02569.x [PubMed: 19046201]

Kaner EF, Beyer F, Dickinson HO, Pienaar E, Campbell F, Schlesinger C, Burnand B. Effectiveness of brief alcohol interventions in primary care populations. Cochrane Database of Systematic Reviews. 2007; 18:CD004148.10.1002/14651858.CD004148.pub3 [PubMed: 17443541]

Kaptchuk TJ, Kelley JM, Conboy LA, Davis RB, Kerr CE, Jacobson EE, Lembo AJ. Components of placebo effect: randomised controlled trial in patients with irritable bowel syndrome. British Medical Journal. 2008; 336(7651):999–1003.10.1136/bmj.39524.439618.25 [PubMed: 18390493]

King M, Nazareth I, Lampe F, Bower P, Chandler M, Morou M, Lai R. Impact of participant and physician intervention preferences on randomized trials: a systematic review. Journal of the American Medical Association. 2005; 293:1089–1099.10.1001/jama.293.9.1089 [PubMed: 15741531]

Kvalem IL, Sundet JM, Rivo KI, Eilertsen DA, Bakketeig LS. The effect of sex education on adolescents' use of condoms: applying the Solomon four-group design. Health Education Quarterly. 1996; 23:34–47. [PubMed: 8822400]

Lipowski SL, Pyc MA, Dunlosky J, Rawson KA. Establishing and explaining the testing effect in free recall for young children. Developmental Psychology. 2014; 50:994–1000.10.1037/a0035202 [PubMed: 24294884]

Mayo, E. The human problems of an industrial civilization. New YorK, NY: MacMillan; 1933.

McCambridge J, Butor-Bhavsar K, Witton J, Elbourne D. Can research assessments themselves cause bias in behaviour change trials? A systematic review of evidence from solomon 4-group studies. PLoS ONE. 2011; 6(10):e25223.10.1371/journal.pone.0025223 [PubMed: 22039407]

McCambridge J, Kypri K. Can simply answering research questions change behaviour? Systematic review and meta analyses of brief alcohol intervention trials. PLoS ONE. 2011; 6(10):e23748.10.1371/journal.pone.0023748 [PubMed: 21998626]

McCambridge J, Kypri K, Elbourne D. In randomization we trust? There are overlooked problems in experimenting with people in behavioral intervention trials. Journal of Clinical Epidemiology. 2014; 67:247–253.10.1016/j.jclinepi.2013.09.004 [PubMed: 24314401]

McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. Journal of Clinical Epidemiology. 2014; 67:267–277.10.1016/j.jclinepi.2013.08.015 [PubMed: 24275499]

Neff RL, Landrum JW. The Life Activities Inventory as a countermeasure for driving while intoxicated. Journal of Studies on Alcohol. 1983; 44:755–769. [PubMed: 6645539]

Parsons HM. What happened at Hawthorne?: New evidence suggests the Hawthorne effect resulted from operant reinforcement contingencies. Science. 1974; 183(4128):922–932.10.1126/science.183.4128.922 [PubMed: 17756742]

Roediger HL, Karpicke JD. Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science. 2006; 17:249–255.10.1111/j.1467-9280.2006.01693.x [PubMed: 16507066]

Sanchez-Craig M, Davila R, Cooper G. A self-help approach for high-risk drinking: effect of an initial assessment. Journal of Consulting and Clinical Psychology. 1996; 64:694–700. [PubMed: 8803359]

Schrimsher GW, Filtz KR. Assessment reactivity: Can assessment of alcohol use during research be an active treatment? Alcoholism Treatment Quarterly. 2011; 29:108–115.

Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and quasi-experimental designs for generalized causal inference. 2. Boston, MA: Houghton Mifflin Co; 2002.

Solomon RL. An extension of control group design. Psychological Bulletin. 1949; 46:137–150. [PubMed: 18116724]

van den Broek GS, Segers E, Takashima A, Verhoeven L. Do testing effects change over time? Insights from immediate and delayed retrieval speed. Memory. 2013; 22:803–812.10.1080/09658211.2013.831455 [PubMed: 23998337]

van Sluijs EM, van Poppel MN, Twisk JW, van Mechelen W. Physical activity measurements affected participants' behavior in a randomized controlled trial. Journal of Clinical Epidemiology. 2006; 59:404–411.10.1016/j.jclinepi.2005.08.016 [PubMed: 16549263]

Weinhardt LS, Carey KB, Carey MP. HIV risk sensitization following a detailed sexual behavior interview: A preliminary investigation. Journal of Behavioral Medicine. 2000; 23:393–398. [PubMed: 10984867]

Willson, VL.; Kim, ES. Pretest sentitization. In: Salkind, NL., editor. Encyclopedia of research design. Thousand Oaks, CA: SAGE Publications, Inc; 2010. p. 1092-1095.

Winch WH. The transfer of improvement in memory in school-children II. British Journal of Psychology. 1908; 2:284–293.

Zelen M. Randomized consent designs for clinical trials: An update. Statistics in Medicine. 1990; 9:645–656. [PubMed: 2218168]

Zgaljardic DJ, Benedict RH. Evaluation of practice effects in language and spatial processing test performance. Applied Neuropsychology. 2001; 8:218–223.10.1207/S15324826AN0804_4 [PubMed: 11989725]

Table 1

Solomon Three- and Four-Group Designs

Design	Randomization	Group Assignment	Baseline/Pre-test	Intervention	Post-test
3-Group	R	Group 1	O_1	X	O_2
	R	Group 2	O_1		O_2
	R	Group 3		X	O_2
4-Group	R	Group 1	O_1	X	O_2
	R	Group 2	O_1		O_2
	R	Group 3		X	O_2
	R	Group 4			O_2

Note. R = Random assignment; O = Observation; X = Intervention.