THE PLACE OF EVALUATIONS IN WORK SAFETY: WHAT CAN WE REALLY ACHIEVE?

HARRY S. SHANNON,

McMaster University, Hamilton, Ontario, Canada

Address:

CRL-221, McMaster University,

1280 Main Street West,

Hamilton, Ontario, L8S 4K1

Canada

Contact for correspondence:

harry.shannon@mcmaster.ca

Word count (text): 4,848

THE PLACE OF EVALUATIONS IN WORK SAFETY: WHAT CAN WE REALLY ACHIEVE?

HARRY S. SHANNON

McMaster University, Hamilton, Ontario, Canada

1. INTRODUCTION:

It is a given that those implementing work safety interventions want those interventions to be effective, that is, they should reduce the risk and rate of occupational injuries. Yet evaluation of those interventions – to demonstrate (or disprove) their effectiveness - has been relatively uncommon (Verbeek, 2007).

There have been numerous 'small' scale measures put into practice over the past hundred or more years. While most interventions were not formally evaluated, many were no doubt effective individually to at least some degree and cumulatively to a large degree. The overall rates of work injury (notably severe and fatal injuries) declined dramatically over the twentieth century, so much so that the Centers for Disease Control in the United States declared the improvements to be one of the 10 great public health achievements of the century (CDC, 1999a).

Yet given the decline in injury rates and the increasing focus on more complex systems, obtaining further substantial improvements will be challenging. Understanding how effective an intervention is - and when and how and for whom - and which interventions are worth disseminating widely requires rigorous evaluation, and as a rule, more rigorous evaluations may be more difficult to accomplish.

When solid evaluations have been done, the results may be disappointing, even if the value of the intervention seemed 'obvious'. For example, Daltroy and colleagues tested an educational program to prevent low back injuries in postal workers. They initially had trouble convincing workplace parties to do a randomized trial. Arguments against the study included "it's not fair to only give the program to some of the workers". But when they did the study, they found no significant difference between the groups. (Daltroy et al., 1997)

While Daltroy's study randomized *individual* postal workers, Hogg-Johnson et al. from the Institute for Work & Health in Toronto, Ontario, randomized 2,153 *companies* with poor safety records to one of three groups. The interventions were part of the Ontario High Risk Initiative, and companies in the groups received one of the two 'treatments' - special consultations or additional inspections - or the control – 'usual service'. The authors checked the reported (workers' compensation) injury rates for nearly two years after the intervention, and found no significant differences between the groups (Hogg-Johnson et al., 2012). There are a number of possible explanations for these results, but the crucial point is that neither intervention worked.

A more disturbing example involved a program designed to improve safety of Afghanis in areas where land mines were prevalent. (The project also studied the impact of a similar program in Angola.) One group of workers -- farmers -- were particularly affected. The overall results were complex, but at least part of the intervention, 'direct training', *seemed to do more harm than good* – there was more risk-taking, leading to more injuries (Andersson et al., 2003).

Of course, these are deliberately chosen examples where the intervention did not work or even made things worse. Other evaluations do show an improvement in safety (see, e.g., CDC, 1999b; Hemenway, 2009: 44-57).

This paper is not intended as a 'how-to' guide. Rather the aim is to discuss the feasibility of doing evaluations of work safety interventions. The criteria to examine when reading work safety evaluations - to decide if they are well conducted and if the results can be trusted - have been described (e.g., Shannon et al., 1999). Likewise, methods for conducting an evaluation are readily available, and have been described specifically for occupational safety interventions (Robson et al., 2001). Those authors noted increasing levels of strength (less susceptibility to bias) for different types of quantitative study. The hierarchy starts with the weakest design, a before-after study, moving to a quasi-experimental approach – involving a control group – and concluding with the scientific ideal, the randomized controlled trial (RCT). As noted earlier, and illustrated by the difficulty of persuading workplace parties to agree to an RCT, more rigorous designs are harder to conduct in practice, especially if the units of randomisation are companies rather than individuals.

2. EXAMPLE OF AN EVALUATION

To illustrate what questions might be answered by good evaluations, I will concentrate in some depth on a particular study. It was very nicely done and covered several relevant issues. I will describe the evaluation and then state some key questions that such evaluations might be expected to answer.

The study looked at the impact of a program implemented by the Dutch Ministry of Social Affairs and Employment. The Ministry offered subsidies to companies to change what could loosely be described as their safety culture. In return, companies had to provide data on injuries before and after the changes, and be open to extra evaluation. The Ministry commissioned 'further research' – the evaluation - which was described in several reports (Hale et al., 2008a; Hale et al., 2008b; Hale et al., 2010).[1] The researchers were given two questions: how to measure "success" in this type of study? And could success as defined be linked to particular interventions, that is, changes made by the companies?

Given that injury rates are susceptible to reporting biases, the authors used a mix of information to determine if any given project was a success. (The 17 'projects' included 29 companies.) The information included trends in injury rates, as well as absences, unsafe behaviours, etc.)

A range of methods, including detailed interviews with key people at each workplace, was used to identify what specific changes each organisation made. The changes were classified in one of three groups: technical changes, organisational changes, or individual and group behavior changes. On average, 16 changes per project were made. The authors labelled each change an 'intervention' (although they might be thought of individually as 'mini-interventions). Examples were: change of director responsible for safety; added inspection rounds and audits; and modification of workplaces, work methods, and good housekeeping campaigns.

The authors then compared the proportion of successful projects which used a given intervention with the proportion of unsuccessful projects that did so. For example, for the intervention "added

[1] Additional material was written in Dutch. I have not read those reports.

inspections and audits", six out of eight (75%) successful projects did this, compared with three out of nine (33%) unsuccessful ones, while the ISO/OHSAS standard was implemented in four (50%) of the successful projects and tree (33%) of the unsuccessful ones. These proportions were compared for all 39 interventions. The authors noted the importance of identifying factors that did not discriminate between the groups. They also divided the lessons that could be learned according to the target audience: the companies, the ministry, and safety scientists.

It is worth noting that the Dutch subsidies study is a rare example of using longitudinal data to analyse work safety across organizations. It investigated how changes made specifically as part of an overall programme were related to improvements in safety performance. This contrasts with a number of reports that correlate injury rates with aspects of a company's safety program like inspections and audits. Several use words like "impact" or "effect" in the titles of their reports (e.g., Arocena et al., 2008; Vredenburgh, 2002). Such language, though, improperly implies that correlation equals causation. Moreover, those studies are almost all cross-sectional, and thus provide at best only weak evidence of causal relationships.

3. FIVE EVALUATION QUESTIONS AND THE FEASIBILITY OF ANSWERING THEM

What questions might have been asked in a study like this? I will examine five, based on the aims of evaluations stated above: understanding how effective an intervention is - and when and how and for whom - and which interventions are worth disseminating. I will discuss if and how the questions were covered in the Dutch subsidies study, who is most interested in which question, and what I see as the general feasibility of addressing them. I will draw on two other studies to supplement the discussion of feasibility.

Question 1: Do the incentives work?

This first question is perhaps the most basic, asking what is the *average* effect? As I noted earlier, the ideal study would be an RCT -- companies might be randomised to receive or not receive subsidies. However, in this case, we could not guarantee that companies would actually make changes. A more feasible study, I think, would be to randomise companies to be offered or not offered subsidies.[2] This reflects a distinction made in clinical epidemiology between an efficacy study (can an intervention work -- in ideal circumstances?) and an effectiveness trial (does it work -- in the real world?). In effectiveness drug trials, for example, patients may not take all their medications, likely diluting the full benefit of the drug. In the Dutch case, the Ministry may offer subsidies, but if very few companies take up the offer, the net (average) effect will be small even if in any given company there is a substantial reduction in risk and injuries. However, that is inherent in the intervention, and so it is an appropriate question to address.

It might be argued that this is treating a very complex intervention – how companies actually implement the changes – as something very simple. This is indeed so, but it is still a legitimate question. A paper 10 years ago was entitled: "Complex interventions: how 'out of control' can a randomized controlled trial be?" (Hawe et al., 2004) RCTs depend on having a standardized intervention, and Hawe and colleagues

[2] In this case, the subsidies were one half of what companies spent on making the changes, so would only have been paid if companies did make changes. However, some companies were apparently already undergoing changes and benefited from the program even though it was not intended for them.

noted that the crucial point depends on just what is standardized. In the example here, would it be the offer of randomization to being offered subsidies, or would it be each mini intervention? Depending on the question being asked, it could be either. Randomising companies to being offered or not offered subsidies allows us to answer the big picture question - what is the average size of effect? The Minister and policymakers are probably most interested in that question, rather than in the specific detail of how companies change their safety culture.

RCTs of individuals or small groups are entirely feasible. One example is Zohar's study of supervisory practices. He was able to randomise supervisors and their work groups to intervention or control group, and observe practices for several periods before, during, and after the intervention (Zohar, 2002). While RCTs that randomise units at a higher level such as companies are more difficult to conduct, they are by no means impossible.[3] One was mentioned above - the Ontario project which randomized companies to one of three groups (Hogg-Johnson et al., 2012). The study was very big involving over 2,000 companies, and was able to use routinely collected data on work injuries rather than having to collect this information directly. It required cooperation between the regulatory agencies – which carried out the consultations or intensive inspections - and the researchers. (This was enabled by a long-standing relationship between the two. The Institute for Work & Health in Toronto is an independent, not-for-profit research organization with core funding from the Province of Ontario.)

Another example of an RCT is from the Netherlands (van der Molen and Frings-Dresen, 2014). At the time of writing, the study is apparently in progress. The aim is to reduce safety violations when working at heights. In May 2014, the protocol was published, this time randomizing *cities* to one of three groups. The groups are to receive face-to-face guidance on hazards, direct mailing of hazard information or no guidance. The study will need 64 companies per group, nearly 200 overall, so it is remarkably ambitious. While the direct mailing of information can of course be standardised, the exact guidance given face-to-face will vary. However, the safety consultants will be divided across cities, so differences in their individual approaches should be evened out across groups. The outcome, a measure of safety violations, will be made by direct observation at each site.

Question 2: For whom does this intervention reduce injury rates?

In the Dutch subsidies program this question was framed as: What are characteristics of successful companies? With numbers too small for statistical testing (eight projects deemed successful, nine unsuccessful), the authors were forced to reach conclusions qualitatively, but were nevertheless able to identify six prerequisites they saw as crucial for success, for example, '[t]he company understands that change is a continuous process …' and '[a]n enthusiastic, creative, active and persistent coordinator is employed or given the time to run the programme' (Hale et al., 2008b).

[3] If single companies make changes at the level of the organization, they would likely do a Before-After comparison. They might also compare the before-after difference in injury rates to the secular trend for their industry.

In a fairly large study (like the Dutch working-at-heights study), and certainly in a very large one like the Ontario study, this question could be answered quantitatively.[4] If a quantitative measure of effect (of the intervention) at the company level is considered valid, a multiple regression with the measure as the outcome can be conducted to establish which company characteristics (independent variables) were most strongly associated with success. If the measure of effect for any company is simply binary – successful or unsuccessful, a multiple logistic regression can be carried out. This sort of information will be most relevant for those in the Ministry who *implement* the subsidies policy, to help them determine where to concentrate their efforts.

Question 3: What components of the overall intervention produce a reduction in injury rates?

Phrased more simply, this asks "*What* works?" What were the 'active ingredients' that the companies applied? If the intervention is complex – which is likely the case in almost any intervention at the organization level – identifying just what was done will need a major effort. This was certainly so in the subsidies study - the authors collected '[v]ery rich data on the interventions themselves' (Hale et al., 2008a), using interviews, documentation and input and process indicators.

As stated earlier, the authors compared the proportions of successful and unsuccessful projects that included each ingredient. In a larger study (again, the Ontario study is an example, although it did not obtain any detail on *what* the companies did in response to the consultations or inspections), this can be done with statistical testing and a greater degree of certainty. The Ontario study authors acknowledged an additional point -- that the way the intervention was implemented might have led to the finding of "no difference". The authors wrote: '[l]imitations in implementation of the method of targeting workplaces, in intensity and duration of programme, and in outcomes available for evaluation may account, in part, for the absence of observed differences among the study groups.'

This is a reminder that the 'intervention' in the study should not be seen as just what the *companies* did. That may have been affected by the quality of the consultations or inspections undertaken. Again, the question of interest depends on who is asking it. The regulatory agencies will want to know what constitutes a high quality inspection or consultation – where 'quality' is defined as what induces workplaces to make positive changes. For workplace parties, i.e., those making changes in the workplace, the germane information is what specific changes to make.

Further questions may still arise about interactions between the specific changes, and whether some changes may be necessary, but not sufficient on their own to affect injury rates. Very large studies might have the statistical power to sort this out, but it is likely to be a question that is difficult to answer, although Hale and colleagues (2008a) noted that the mix of actions in the two most successful companies were very similar.

Question 4: What are the mechanisms for success?

[4] The Ontario study had information on the 'rate group' (type of industry), age of the company, number of firm branch locations, geographic region, and firm size.

This 'how?' question aims to provide the link between the components of the intervention and the outcome. In many evaluations, researchers construct a 'logic model' that shows *how* the process is expected to unfold. A simple example for an educational programme to prevent back injuries is shown in Figure 1. The material has to be appropriate, workers have to access the material, and understand it, the opportunity to use the techniques taught must be present, and workers have to comply. The techniques themselves must reduce back strain, which in turn should reduce injuries. To the extent any of these do not apply - for example, few workers access the training - the effect of the programme will be reduced or even nullified.

Some of these issues might be answered quantitatively, but a full understanding almost certainly requires qualitative methods, often via interviews with people in the workplace, observations of the workplace, reviews of documents, and/or focus groups. The aim is to understand what happened and *how* did it happen. In the subsidies study, these questions were intended to identify differences between the successful and unsuccessful projects, and the authors defined 12 steps in designing and conducting the interventions (Hale et al., 2008b)

While Hale and colleagues listed these 12 steps under 'Lessons for companies', they will also be of crucial importance for those who develop interventions. They will have two main issues -- did we do the right thing or are there changes we can propose in implementation? They will also want to learn if the underlying theory (which links to the logic model) is applicable. As the social psychologist Kurt Lewin famously wrote: 'nothing is as practical as a good theory' (Lewin, 1945: 129). This apparently paradoxical remark makes the point that if the theory is good, it will help in developing appropriate interventions, since it will allow some degree of generalizability in developing useful actions.

Question 5: What is the cost-benefit ratio?

Are we using resources well? Or, more loosely, is the intervention worth the effort and expense? (I am ignoring any ethical or legal obligations to reduce injuries.) In principle, the methodology for this economic analysis is straightforward. First we need to identify costs such as time needed by staff to implement the intervention and how much is spent on safety. The reductions in costs of insurance premiums, etc, can be subtracted to obtain the net costs. The answers to the earlier questions provide estimates of the reductions in injuries, and these benefits are compared with the net costs. In practice, obtaining the cost estimates may not be simple, although it should be feasible.

As well, economists point out that 'costs' depend on whose perspective one takes. The viewpoints of workers, companies, workers' compensation systems, and societies, and hence their calculations of the worth of a program, will all differ. However, provided the perspective is clearly specified, the corresponding analysis can be done and appropriate conclusions drawn.

4. DISCUSSION

The text above is optimistic about the possibility of answering the key questions about workplace interventions – do they work, how strongly, for whom, under what circumstances? - with a good if not

high degree of rigour. Some will say it is too optimistic – indeed having completed the Dutch subsidies study, those authors (Hale and colleagues) made several statements on the nature of interventions and the feasibility of doing evaluations that showed some doubts. I will examine four of their key statements.

1. "Changes in workplaces to improve safety do not come in single, neat packages …" I agree and note that the text on question 3 above referred to the different sets of interventions made, what I called the specific ingredients that comprised the overall package of changes implemented by the companies. I also argued that if the question asks if the overall approach (offering subsidies) is the question of interest, the complexity of the intervention does not matter.

2. "Evaluation stands or falls with the existence of quantitative indicators of safety in companies". Again, I agree. Qualitative methods can give a sense of *whether* an intervention is effective, but cannot quantify the degree of success. Thus, if qualitative studies found that Intervention A compared with normal practice was effective, and likewise Intervention B was better than normal practice, one could not specify which of the two interventions was superior. That would require quantitative measures of the outcome.

Yet the problem of accurate measures has long been recognized and remains. Many authors have written about this (e.g., Shannon and Manning, 1979). Indeed, more than 40 years ago, Hale and Hale acknowledged that injuries might not be correctly reported, especially after an intervention. They noted that it might not be the number of *accidents* that is suppressed, but rather the number *reported*, a concern still repeated (e.g., Brown and Barab, 2007; Lipscomb et al., 2013), and they concluded: "[t]*he most satisfactory procedure would seem to be the use of as many criteria [measures] as possible in evaluating the success of a safety program."* Hale and colleagues did this in the Dutch subsidies study.

An alternative to using a 'lagging' indicator, number of injuries, is to measure the level of 'safety' – or some facet of safety - in a workplace, a 'leading' indicator. Such a measure, based on observations of pertinent safety violations, is the outcome in the multi-city study of falls prevention (van der Molen and Frings-Dresen, 2014). This obviates the concern about reporting biases, but it can still beg the question: does the leading indicator truly reflect the level of safety? "Validating' it against the number of injuries would constitute circular argument. Likewise, measures intermediate between the intervention and the 'true' outcome of injuries might be used, e.g., safety behavior following an intervention. They, too, suffer from the validation problem.

Still, I think that using several measures can be valuable. Hale and colleagues did so in their study of subsidies, pooling all the information to classify the 17 projects as successful or unsuccessful – a binary measure that did not explicitly quantify the degree of success.[5] An alternative would have been to use separately several quantitative measures, and examine the agreement among them in determining whether the level of safety for each measure had improved, and calculating the size of any change comparing projects that applied or did not apply an intervention.

I expect that attempts to develop valid and reliable quantitative measures will continue. For example, Amick and colleagues have developed an eight item scale for workplaces (the IWH Organizational Performance Scale) that predicts future injury rates, albeit with a limited degree of accuracy so far

[5] They did identify two projects that they labelled 'most successful'.

(Amick, personal communication). Whether we can do better (and be sure we have done better) than using 'as many criteria as possible', as recommended decades ago, remains to be seen. For small businesses, that may rarely if ever experience a serious accident, evaluation must rest on some measure other than injury rate or on data pooled across a number of such organizations.

3. "Strictly controlled scientific methods of epidemiology can almost never be used …" While there are often barriers to conducting randomized trials, some have been done. As I noted earlier, there are a good number of examples of RCTs of individuals or work groups. As well, there is an increasing number of RCTs of companies, and even one under way that has randomized cities. To be sure, they rely on special circumstances, and perhaps good fortune for the investigators that opportunities to do evaluations occur. RCTs could also become more common if researchers form links with workplace parties and policy makers, so the latter are primed to conduct high quality evaluations of interventions. The priming should allow evaluations to be as rigorous as possible, e.g., through planning *before* the interventions are implemented.

There may also be ethical limitations. RCTs should normally be done only if there is 'equipoise', that is, the investigators are genuinely unsure about whether the new approach is better or worse or no different than the old one. However, RCTs are also ethical if there are limited resources, for example, only some companies can receive inspections or consultations. If so, one can argue that the fairest way to determine which companies are in the 'treatment' and 'control' groups is through randomization.

4. "The methodology was based on the highest level of scientific measurement available, namely a before and after study of the trends in performance indicators." It appeared that the request for evaluation of the Dutch subsidies program was made by the Ministry after the program was well under way. Given that, the comment is not surprising. For example, there was no opportunity for including a comparison group. As I noted above, I believe an RCT that compared offering and not offering companies the subsidies would have been feasible, and thus would have been a higher quality study.[6]

4.1 Other challenges to evaluation

There are of course other challenges to doing evaluation. One important one is creating a culture of evaluation. I touched on this above, in noting that if plans for evaluation had been made earlier in the Dutch subsidies study, a more rigorous evaluation could have been conducted. The term 'evidence-based' has found its way into many public and scientific discourses. This provides us with a lever to remind all interested parties that we cannot be evidence-based if we do not develop the relevant evidence! And this requires proper evaluations. Since one barrier can be the lack of ability to do the studies, more partnerships between 'expert' evaluators and workplaces may help, especially if the evaluators pass on their expertise, allowing workplaces to become self-reliant in doing evaluations.

[6] The authors stated that they did not have a comparison group. This was true for the overall question on whether subsidies are effective in reducing companies' injury rates. For each specific intervention, the authors compared the proportion of successful and unsuccessful companies that implemented the intervention. Epidemiologists will recognize this as analogous to a case-control approach. I believe the authors could have instead treated the data using a cohort model. The companies that did and did not apply the intervention formed two groups, and for each the proportion that were successful or not could have been calculated. The proportions in the two groups could then have been compared.

Persuading companies to participate in evaluations is not always easy. If a safety manager has staked her reputation on promoting new procedures or interventions, she may be reluctant to learn that it was all for nothing. Again, creating a culture of evaluation may help this, and government agencies could be encouraged to provide incentives for companies to participate. We should also take advantage of opportunities that occur, that do not require difficult-to-obtain cooperation. Levine et al. (2012) studied the impact of government safety inspections of high-injury companies in California, among which the ones inspected were chosen randomly. The authors matched each of them with a high-injury company not inspected, and compared the two groups not only on injuries over the next four years, but on other variables as well.

A further challenge is to apply what we already know, while making sure that the safety measure really does work in a new context. Would a policy that is effective in, for example, Sweden have a similar impact in Asia, or in Africa, or in the United States? Developing and testing new or adapted approaches in lower income countries is crucial for work safety on a global level. Indeed, replication is a basic principle of science – and even when applying a 'proven' strategy in a similar context, confirming its impact on safety is important. Social media are creating new ways to present old information, and we need to determining if the impact is improved.

In discussing Question 4 above, I noted that qualitative methods were crucial. In general, I think that most interventions should include qualitative methods -- or at the very least seriously consider whether they can help. For example, an evaluation could conduct a big-picture analysis of a large sample, and a more detailed look at a smaller targeted sample. Combining quantitative and qualitative approaches, known as 'mixed methods', is now much more common than even in the recent past.

5. SUMMARY

To summarise: many interventions have been implemented, but few evaluated. Even so, work safety has improved substantially over time. Nevertheless, when evaluations are done, some interventions are found to have little or no effect, or even make things worse. Given the relatively low injury rates today, further improvements may well need guidance from good evaluations. Of course, different stakeholders are interested in different questions, and we need to tailor the methods to the question being asked. The questions asked should be important ones; there is no point doing a full-blown RCT of minor interventions. And we must grapple with new methodologies to answer the big questions. With foresight and effort, evaluations that have at least a good degree of rigour are often feasible, and more evaluations should be conducted. This will allow safety practitioners to make and justify assertions that their practices are evidence-based.

Note: This paper is based on a keynote talk at the WoSNet.2014 Working on Safety in Scotland, October 2014.

I am grateful to Andrew Hale and his colleagues for their help in providing me with additional material and their scientific integrity in encouraging me to critique their work.

REFERENCES

Andersson N Swaminathan A, Whitaker C, Roche M (2003). Mine smartness and the community voice in mine-risk education: lessons from Afghanistan and Angola. Third World Quarterly 24: 873-887.

Arocena P, Núñez I, Villanueva M (2008). The impact of prevention measures and organizational factors on occupational injuries. Safety Science 46:1369-1384.

Brown GD, Barab J (2007). "Cooking the books" – Behavior-based safety at the San Francisco Bay Bridge. New Solutions 17(4):311-324.

CDC (1999a). Ten Great Public Health Achievements -- United States, 1900-1999**.** MMWR 48(12);241-243.

CDC (1999b). Achievements in Public Health, 1900-1999: Improvements in workplace safety--United States, 1900-1999. MMWR 1999; 48:461-9.

Daltroy LH, Iversen MD, Larson MG, Lew R, Wright E, Ryan J, Zwerling C, Fossel AH, Liang MH (1997) A Controlled Trial of an Educational Program to Prevent Low Back Injuries. New England Journal of Medicine 337:322-328.

Hale AR, Guldenmund FW, van Loenhout PLCH, Oh JIH, Evaluating safety management and culture interventions to improve safety: Effective intervention strategies. Safety Science 2010; 48: 1026-1035.

Hale A, Guldenmund F, Oh J, van Loenhout P, Booster P, Oor M (2008a). Evaluating the success of safety culture interventions. In: Proceedings of the International Congress on Probabilistic Safety Assessment & Management. Hong Kong.

Hale AR, Guldenmund FW, van Loenhout PLCH, Oh JIH (2008b). Evaluating safety management and culture interventions to improve safety: effective intervention strategies & lessons learned. Paper to the Working in Safety Network Conference, Crete.

Hale AR, Hale M (1972). A review of the industrial accident research literature. Committee on Safety and Health at Work Research Paper. London: Her Majesty's Stationery Office.

Hawe P, Shiell A, Riley T (2004). Complex interventions: how "out of control" can a randomised controlled trial be? BMJ 328:1561-1563.

Hemenway D (2009). While we were sleeping: success stories in injury and violence prevention. Berkeley, CA: University of California Press.

Hogg-Johnson S, Robson L, Cole DC, Amick BC III, Tompa E, Smith PM, van Eerd D, Mustard C. A randomised controlled study to evaluate the effectiveness of targeted occupational health and safety consultation or inspection in Ontario manufacturing workplaces. Occup Environ Med, 2012; 69: 890-900.

Levine DI, Toffel MW, Johnson MS (2012). Randomized government safety inspections reduce worker injuries with no detectable job loss. Science 336: 907-911.

Lewin K (1945). The Research Center for Group Dynamics at Massachusetts Institute of Technology. Sociometrics 8: 126-136.

Robson LS, Shannon HS, Goldenhar LM, Hale AR (2001). Guide to Evaluating the Effectiveness of Strategies for Preventing Work Injuries: How to Show Whether a Safety Intervention Really Works. DHHS (NIOSH) Publication No. 2001-119. Cincinnati, Ohio: NIOSH.

Shannon HS, Manning DP (1979). A note on reported accident rates. J Occup Accidents 2:245-253.

Shannon HS, Robson LS, Guastello SJ (1999). Methodological criteria for evaluating occupational safety intervention research. Safety Science 31: 161-179.

van der Molen HF, Frings-Dresen MHW. Strategies to reduce safety violations for working from heights in construction companies: study protocol for a randomized controlled trial**.** BMC Public Health, 2014; 14: 541.

Verbeek J (2007). Cochrane Corner: Occupational injuries. Injury Prevention 13:13-14.

Vredenburgh A (2002). Organizational safety: Which management practices are more effective in reducing employee injury rates? J Safety Res 33:259-276.

Zohar D (2002). Modifying Supervisory Practices to Improve Subunit Safety: A Leadership-Based Intervention Model. J Appl Psychol 87:156-163.