Discussion paper

Scand J Work Environ Health 2015;41(5):491-503

doi:10.5271/sjweh.3505

Evaluation of occupational health interventions using a randomized controlled trial: challenges and alternative research designs

by Schelvis RMC, Oude Hengel KM, Burdorf A, Blatter BM, Strijk JE, van der Beek AJ

This overview aims to guide researchers in occupational health in conducting evaluative research. Several appropriate alternatives for the randomized controlled trial design are available and feasible (ie, stepped wedge, propensity scores, instrumental variables, multiple baseline design, interrupted time series, difference-in-difference, and regression discontinuity), which may provide sufficiently strong evidence to guide decisions on implementation of interventions in workplaces.

Affiliation: Netherlands Organization for Applied Scientific Research TNO, Schipholweg 77-89, 2316 ZL Leiden, The Netherlands. karen.oudehengel@tno.nl

Refers to the following texts of the Journal: 1999;25(6):589-596 2010;36(1):25-33 2012;38(1):27-37 2013;39(1):57-65

The following articles refer to this text: 2015;41(5):421-424; 2016;42(4):273-279; 2016;42(4):257-259; 2019;45(2):194-202; 2019;45(3):318-319

Key terms: difference-in-difference; effectiveness; evaluation; experimental study design; instrumental variable; interrupted time series; multiple baseline design; observational study design; occupational health intervention; propensity score; randomized controlled trial; RCT; regression discontinuity; research design; stepped-wedge design

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/26030719

Discussion paper

Scand J Work Environ Health. 2015;41(5):491–503. doi:10.5271/sjweh.3505

Evaluation of occupational health interventions using a randomized controlled trial: challenges and alternative research designs

by Roosmarijn MC Schelvis, MSc,[1, 2, 3, 4] Karen M Oude Hengel, PhD,[1, 2, 3] Alex Burdorf, PhD,[5] Birgitte M Blatter, PhD,[2, 3] Jorien E Strijk, PhD,[2] Allard J van der Beek, PhD [3, 4]

Schelvis RMC, Oude Hengel KM, Burdorf A, Blatter BM, Strijk JE, van der Beek AJ. Evaluation of occupational health interventions using a randomized controlled trial: challenges and alternative research designs. *Scand J Work Environ Health.* 2015;41(5):491–503. doi:10.5271/sjweh.3505

Occupational health researchers regularly conduct evaluative intervention research for which a randomized controlled trial (RCT) may not be the most appropriate design (eg, effects of policy measures, organizational interventions on work schedules). This article demonstrates the appropriateness of alternative designs for the evaluation of occupational health interventions, which permit causal inferences, formulated along two study design approaches: experimental (stepped-wedge) and observational (propensity scores, instrumental variables, multiple baseline design, interrupted time series, difference-in-difference, and regression discontinuity). For each design, the unique characteristics are presented including the advantages and disadvantages compared to the RCT, illustrated by empirical examples in occupational health. This overview shows that several appropriate alternatives for the RCT design are feasible *and* available, which may provide sufficiently strong evidence to guide decisions on implementation of interventions in workplaces. Researchers are encouraged to continue exploring these designs and thus contribute to evidence-based occupational health.

Key terms difference-in-difference; effectiveness; experimental study design; instrumental variable; interrupted time series; multiple baseline design; observational study design; propensity score; RCT; regression discontinuity; stepped-wedge design.

The randomized controlled trial (RCT) is considered the gold standard in evaluative medical research as causal inferences about the therapy under study can be drawn. The first RCT was reported in a 1948 issue of the British Medical Journal (BMJ) and involved the experimental treatment of pulmonary tuberculosis (1). In this trial, a particular group of English tuberculosis patients from different care facilities, comparable in the symptoms of the disease and age were included. The included patients were assigned to either a combined medicine and bed-rest therapy, or bed-rest therapy alone, based on a statistical series of random sampled numbers. Neither the patients nor the doctors involved knew the condition the patient was assigned to, later to be named a "double blind" procedure. Therapy progress was reported on forms particularly designed for this trial. Due to this design, the researchers were able to demonstrate the added value of the combined treatment over the bed-rest treatment, but only in the first three months after onset of the disease. Thereafter a deterioration emerged, probably due to resistance to the medicine under study. Many researchers have followed this example ever since. The beauty of the randomization procedure is that chance (probably) ensures that known and unknown prognostic factors are balanced over the treatment conditions and thus do not interfere with the treatment–outcome relationship. Therefore, conclusive statements about the effectiveness of the therapy can be made.

1 Both authors contributed equally to this work.
2 Netherlands Organization for Applied Scientific Research TNO, Leiden, The Netherlands.
3 Body@Work, Research Center on Physical Activity, Work and Health, TNO-VU/VUmc, The Netherlands.
4 Department of Public and Occupational Health, The EMGO+ Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands.
5 Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands.

Correspondence to: KM Oude Hengel, Netherlands Organization for Applied Scientific Research TNO, Schipholweg 77-89, 2316 ZL Leiden, The Netherlands. [E-mail: karen.oudehengel@tno.nl]

In occupational health research, a typical RCT aims, for instance, to reduce productivity loss at work (ie, a primary outcome) for a randomly chosen group of employees with medically verified upper-extremity disorder (ie, specific characteristics) via an ergonomic assessment at the worksite and a physician contacting each employee's supervisor to discuss potential accommodations at work (ie, a multicomponent intervention). The effectiveness of the intervention is evaluated by the change in primary outcome from pre- to post-test in the intervention group relative to the change in this outcome in the reference group that did not receive the intervention (2). However, occupational health researchers are increasingly addressing questions regarding the outcomes of complex interventions. A complex intervention can consist of (i) multiple components, (ii) multiple providers and thus multiple levels, (iii) multiple locations, and/or (iv) multiple (varying) outcomes. The components, providers, locations and outcomes are interdependent and therefore the intervention can be difficult to standardize or administer uniformly (3–5). Furthermore, the context is often complex and thus nearly impossible to control entirely (6). Conducting an RCT on a complex intervention within an occupational health context is thus not always the most feasible option (7, 8).

The British Medical Research Council (MRC) recently published an updated guide that underlines the need for innovative evaluation methods (9). Although the MRC considers individual randomization in trials as the most robust design to prevent allocation bias, it is more and more acknowledged that common evaluation methods are not always practical or ethical for complex interventions (9). The RCT sometimes even offers too little information to draw meaningful conclusions for science or practice. More specifically, an RCT allows conclusions on the effectiveness of the intervention for a selected sample of individuals. Researchers have argued that because of complexity in the intervention and context, the required conditions that are needed for an efficacy trial will never occur (10). Even if an efficacy trial has been performed with success, then it still is "highly unlikely that interventions that do well in efficacy studies will do well in effectiveness studies, or in real-world applications" [(10) p1262].

In order to further develop the evidence base in occupational health there is a clear need for alternatives to the RCT. These alternatives can be formulated along two lines: experimental (most often RCT variants) and observational studies (11). Some (experimental) alternatives have been applied already in the occupational setting. The most commonly applied RCT variant is the cluster RCT, in which groups of individuals rather than individuals are randomized (12). Cluster RCT typically involve two levels, the cluster (eg, department) and their individual members (eg, worker), although trials of more than two levels (eg, company, department, and worker) also exist (12). Cluster RCT have several advantages over individual RCT in organizational interventions, namely (i) increased logistic feasibility in delivering the intervention, (ii) analysis and evaluation is conducted at the same level as the intervention is applied to (ie, the group), and (iii) contamination is avoided, which might occur when unblinded interventions are administered to some individuals but not to others in the same setting (eg, department, team, occupational physician) (13). Another commonly applied variant is the controlled trial wherein a selected group of individuals or clusters receiving the intervention is compared to a reference group that is matched on known prognostic factors (eg, age) (14). This design can be preferable to an individual or cluster RCT for practical or ethical reasons in an occupational setting. Apart from randomization, the controlled trial shares all characteristics with an RCT, but lacks the advantage of balanced unknown prognostic factors in both conditions.

However, for these alternative RCT designs, challenges remain that impede drawing causal inferences (15). The cluster RCT, for example, needs much larger numbers of participants within an experimental setting, which is often problematic in terms of feasibility and costs. The controlled trial suffers from the non-random allocation to groups, which may introduce known and unknown factors to be unbalanced between both groups. This article presents an overview of other experimental and observational study designs for occupational health interventions, starting with an overview of practical challenges in conducting an RCT, the methodological consequences of these challenges, and an empirical example. Thereafter, the key features of each design are described, including the advantages and disadvantages, and how the challenges are minimized by applying this design.

Challenges in applying RCT to evaluate complex interventions

When conducting an RCT in the occupational setting, researchers faces challenges concerning the methodology (ie, randomization and control group), the intervention, and the context. Empirical examples for each challenge are given in table 1.

Methodology: randomization

Randomization of participants to the experimental condition (intervention group) or usual care/placebo condition (control group) eliminates allocation bias and internal validity threats, maximizing the probability that (un)known confounding variables will be evenly distributed over groups (16).

Table 1. Challenges, their examples, and consequences in occupational health intervention studies.

Clustered main challenge	Challenge	Example [a]	Consequence
Randomization	Only few clusters at organizational level are available to evaluate the intervention	A case management approach for workers on prolonged sick leave was evaluated in one hospital compared to a neighboring hospital (74).	Potential confounding due to differences between groups resulting from unreliable randomization
	The organization objects to random allocation to an intervention or control group	Department supervisors allocated participants to a prevention program, which could have reflected their subjective interest in the program, the time and workload to complete the program, and their expectations for enhancing workplace mental health (75).	Potential confounding due to differences between groups resulting from selection bias
Control group	The organization wants to target all employees with an intervention	An intervention on behavioral techniques was implemented in two departments within a hospital. However, the content of the intervention was different between the two departments (76).	Difficult to differentiate effects of the intervention from changes due to other causes resulting from the lack of a control group
Intervention	The organization or researcher wants to adjust the intervention protocol	The goal of the intervention was to increase the range of healthy foods on offer in worksite cafeterias in two supermarkets in The Netherlands so as to discourage eating snacks. There were conflicting interests between the intervention's goal and management's targets: the greatest profits came from selling snacks (77).	Difficult to differentiate which effects were caused by which element of the intervention resulting from changes in the intervention protocol
Context	The organization is subject to internal change	An intervention to improve sustainable employability was conducted across multiple companies and worksites. During the intervention period, workers from several worksites were discharged as result of the economic recession (78).	Potential confounding due to differences between groups resulting from selective loss to follow-up
	The organization is subject to external changes	In a construction company in The Netherlands, an intervention is introduced to decrease inhalation of particulate matter at the worksite. At the same time, the Labor Inspectorate decides to enforce regulations and plans to visit all worksites in the Netherlands (79).	Difficult to differentiate effects of the intervention from those due to unintended co-interventions resulting from external changes

[a] The empirical examples were selected on relevance and appropriateness for the occupational setting.

Challenge 1. Only few clusters at the organizational level are available to evaluate the intervention. Many workplace interventions are implemented at the group level (eg, company, facility, department, and team). The randomization procedure is then applied at the group or – in methodological terms – cluster level. However, recruiting enough clusters within a specific context is often difficult. If too few clusters are included, controlling by chance for all factors and conditions that might differ between groups is impeded. Consequently, there might be an unequal distribution of baseline characteristics between groups which introduces bias to the study [eg, (17)].

Challenge 2. The organization objects to random assignment of persons or departments. In practice, acknowledgment of a problem which is unique to a certain department (eg, high sickness absence, lagging work performance) can be a strong driver for organizations to participate in intervention research. Targeting this department with an intervention is at least in their interest and at the most a precondition to participate. Thereby, companies obstruct randomization and potential bias is thus introduced. If the organization wants to decide on the allocation of the employees to the intervention and control group, the two most important resulting biases are confounding (ie, error due to a third variable that influences the exposure–outcome relation) (16) and selection bias (ie, error due to systematic differences in characteristics such as motivation, between intervention and control group) (16), which are difficult to overcome (18).

Methodology: control group

The effect of an intervention is measured as the difference on a certain outcome between the intervention group and the control group (18). A control group is needed to distinguish between change in outcome over time due to the planned intervention, or changes over time due to unmeasured or unknown factors (eg, a policy measure).

Challenge 3. The organization wants to target all employees with an intervention. Organizations are often willing to participate in an intervention study if an acknowledged problem is to be solved (eg, high prevalence of low-back pain). Hence, the employer is motivated to demonstrate that (s)he takes the problem seriously and therefore demands that everyone should be able to participate. The employer considers it unethical to offer the intervention to a selected group only, while every employee has a potentially elevated risk. As a consequence, studies within the occupational health setting sometimes have to be performed without a control

group, complicating the distinction between effect of the intervention and autonomous change over time.

Intervention

When following the guidelines of conducting an RCT, a predefined protocol for implementing and evaluating interventions is preferred in order to reach high internal validity (16). High fidelity to the protocol is furthermore important in order to understand key intervention processes and functions, and thus enable answers to the question of why the intervention is or is not effective.

Challenge 4. The organization or the researcher wants to adjust the intervention protocol. Either the organization or the researchers may want to adjust the intervention protocol to fit the specific context per cluster, thereby violating the standardization principle. For instance, the order of intervention components may be altered or the intervention components may be tailored to a specific group of workers or to specific occupational health problems. If adjustments are made within clusters, it becomes difficult to establish which intervention components or what implementation processes contributed to the effectiveness or lack thereof of the intervention, a situation sometimes referred to as a 'black box' (19, 20).

Context

For most occupational health interventions a double-blind-placebo trial is nearly impossible: complex interventions are dependent on the context in which they are applied. Moreover, besides the intentional adjustments described under challenge four, the intervention provider, the participants or the context may unintentionally influence the delivery and content of the intervention and thereby the outcomes (ie, information bias).

Challenge 5. The organization is subject to internal change. Many worksites and departments are subject to continuous change (21). For example, within the participating department not only the intervention under study, but also a co-intervention is delivered. In hospitals, lifting devices may be introduced to reduce mechanical load among nurses, whereas simultaneously hospital beds are replaced by high-low beds that also reduce nurses' mechanical load. Implementing the intervention under fully controlled conditions is thereby impeded. A second example is a change in staffing: employees and managers change jobs or retire, new employees are hired, and teams are moved to other areas or downsized. Consequently, high loss to follow-up can be expected and a decreased study sample complicates reliable conclusions regarding intervention effects.

Challenge 6. The organization is subject to external change. Even when the intervention is performed under controlled conditions within the company, external changes might interfere with the intervention (21). For instance, increased enforcement of regulations by the Labour Inspectorate on the main outcome of the intervention might take place simultaneously (eg, exposure to dust containing quartz). Or, a nationwide campaign on work stress is implemented during the same period as a local stress management intervention, which motivates the control group to implement stress prevention measures as well. As a consequence of these so-called co-interventions, it becomes more difficult to distinguish autonomous change from effect, even if a control group is present.

In sum, difficulties with regard to methodology, intervention, and context may hamper the evaluation of complex occupational health interventions by means of an RCT. However, we fully agree with Kristensen (18), who stated that "there may be many good reasons for not performing a RCT in an occupational setting. But there are no good reasons for ignoring the problems created by not applying such a design."

Alternatives for evaluating complex occupational health interventions

Several alternative experimental designs and designs using observational data are potentially interesting for the evaluation of complex interventions (14, 22). The core team of authors discussed a list of potential alternatives for the occupational health setting and those most relevant and applicable to the occupational health setting were selected for this article. In contribution to the current debate on alternatives to randomization in the evaluation of public health interventions (9), the selection of alternatives is described based on theoretical literature and empirical examples (tables 2 and 3).

Alternative design in experimental research

Stepped-wedge randomized trial. The stepped wedge randomized trial is a modification of the individual or cluster RCT in which an intervention is sequentially rolled-out to all participants over consecutive time periods (23). The order in which the individuals or clusters receive the intervention is randomized, so that at the end of the entire time period all participants have received the intervention, thereby counteracting challenge 3 (*the organization wants to target all employees with an intervention*) (24). The stepped-wedge design is particularly suitable if it is considered unethical to withhold the intervention from participants in a control group (25). Additionally, the stepped wedge design allows for improvement of the intervention based on

Table 2. Overview of evaluative designs, their characteristics, and data requirements. [RCT=randomized controlled trial.]

Design type	Allocation of intervention	Confounding	Data requirements	Measurements
Experimental				
RCT	Randomization at individual level	Known and unknown prognostic factors are balanced	Longitudinal data	Before and at least once after intervention
Cluster RCT	Randomization at group level	Known and unknown prognostic factors may be unbalanced over clusters	Longitudinal data	Before and at least once after intervention
Stepped-wedge randomized trial	Randomization of intervention to all individuals or groups sequentially over time	Known and unknown prognostic factors may be unbalanced over clusters and over time	Longitudinal data	Repeated before and after
Observational				
Propensity score method	Likelihood to have been offered the intervention	Matching, stratification or adjustment for known prognostic factors	Longitudinal data	Before and at least once after intervention
Instrumental variable method	Exposure to 'the instrument' predicts actual intervention received	No influence of unknown prognostic factors	Longitudinal data	Before and at least once after intervention
Multiple baseline design	Intervention to all individuals or groups sequentially over time	Adjustment for known prognostic factors	Longitudinal data	Repeated before and after
Interrupted time series	Intervention to all individuals at particular moment in time	Adjustment for known prognostic factors	Cross-sectional data	Multiple repeats over time (eg, routinely collected data) before and after
Differences-in-differences	Intervention to selected individuals at particular moment in time	Adjustment for known prognostic factors	Cross-sectional data	Multiple repeats over time (eg, routinely collected data) before and after
Regression discontinuity	Intervention to individuals at particular moment in time	No influence of unknown prognostic factors	Cross-sectional data	Multiple repeats over time (eg, routinely collected data) before and after

lessons learned in every subsequent step (which makes it very suitable for effectiveness trials in practice) and thereby eliminates challenge 4 (*the organization or the researcher wants to adjust the intervention protocol)*. Due to the within and between cluster comparisons at each measurement time across all time periods, this design allows for a variety of conclusions: both short- and long-term effects, fade out effects, and the natural course of the condition under study (26).

For the evaluation of a care program for staff members in dementia special care units, a stepped-wedge design was used (27). The care program consisted of tools and procedures to guide staff members through the detection, analysis, treatment and evaluation of residents' challenging behavior. After allocating seventeen units randomly to five groups, every four months a new group started with the intervention (24 months in total). Burnout, job satisfaction, and job demands were self-assessed before the start, midway and after the implementation process. The results of the multilevel analyses of 380 staff members showed a significant positive effect for job satisfaction [β 0.93, 95% confidence interval (95% CI) 0.48–1.38], whereas no statistically significant effects were found for burnout and job demands.

Although the stepped-wedge design helps to minimize or overcome two important challenges, it introduces new challenges in itself. These challenges are firstly that larger sample sizes might be required for some outcomes since, with the increased number of groups to compare, the design may have less statistical power than the regular (cluster) RCT (28, 29). Secondly, the data collection in each time period can put a high burden on participants and researchers, which might hamper the feasibility of the study (29). The design is most feasible if data can be (partly) routinely collected at the appropriate time intervals in a reliable and valid way (28). Thirdly, statistical analysis is complex because both a random coefficient for cluster and a fixed effect for time need to be taken into account (23).

Alternative designs in observational studies

In observational studies, assignment to the experimental condition is not under the researchers' control. The intervention and control group may differ in (observed) covariates, which could lead to biased estimates of intervention effects. Hereafter we describe alternative evaluation designs specifically developed to evaluate interventions with observational data while dealing with potential bias.

Propensity scores

The propensity scores method is a statistical matching technique that can be applied to control for confounding in evaluation studies with observational data (30, 31). The first step is to estimate propensity scores for all individuals, defined as the conditional probability of (a par-

Table 3. Alternative research designs, advantages and disadvantages, and the challenges overcome. [RCT=randomized controlled trial.]

Alternatives	Advantages	Disadvantages	Solution to challenges [a]
Stepped-wedge design	All participants and clusters receive the intervention Randomization of order in receiving the intervention, thus preventing selection bias Improves feasibility for practical, ethical and/or financial reasons Changes in protocol are possible before the next step	More measurements and a longer time period High burden on participants and researchers Complexity of statistical analyses Decrease in statistical power and requires larger sample sizes	3, 4
Propensity scores	Solution if randomization has unethical consequences Advantageous when effects are expected in the long term, which makes an RCT costly	Very large sample sizes are needed If the propensity score is estimated incorrectly or the covariates are measured imperfectly, bias is introduced	2
The method of instrumental variables	Solution if an RCT is not possible for practical reasons Instrumental variables can also correct for unmeasured confounders	Rarely used in research practice because of strong assumptions Weak instruments lead to large standard errors	2
Multiple baseline design	Fewer cohorts are required as cohorts act as their own controls Analyses can be done with routinely collected data, cohort data or (historic) reference groups	Including fewer cohorts is only possible if the effect sizes are large and if the intervention results in rapidly observable changes in the outcome variable The autocorrelation can lead to inaccurate estimates of the intervention effect Sufficient baseline stability is needed The timing of the intervention introduction and duration in each cohort ought to be known before starting	1, 3, 5, 6
Interrupted time series design	Suitable if establishing control groups is difficult Analyses can be based on routinely collected data, cohort data or external (historic) data or reference groups	Underpowered studies because of the number of measurements before and after the intervention and the time lags of measurements One needs to be able to determine specifically at what time point the intervention started and ended	2
Differences-in-differences	An elegant way to study 'naturally occurring' internal or external changes	Sophisticated analyses required Does not account for invariant factors or macro trends that interfere with the outcome Under- or overestimation is a risk in individual level interventions	5, 6
Regression discontinuity	Establish causal effects based on observational data Feasibility of this design can be improved by using routine clinical or administrative data	Assumption that individuals around the threshold are similar is debatable The assignment variable possibly changes over time Requires larger samples sizes than an RCT	1, 2, 3

[a] Challenge 1: Only few clusters exist to cluster the intervention at organizational level; Challenge 2: The organization objects to random assignment of persons or departments; Challenge 3: The organization wants to target all employees with an intervention; Challenge 4: The organization or the researcher wants to adjust the intervention protocol; Challenge 5: The organization is subject to internal change; Challenge 6: The organization is subject to external change.

ticular) exposure to the intervention given a number of confounding variables (32). The propensity score can be estimated with logistic regression analysis, modeling the exposure as dependent variable and the potential confounders as independent variables (33). Because some individuals with similar propensity scores are exposed to the intervention, whereas others with a similar score are not, the method assumes that actual exposure to the intervention within these individuals mimics randomization (34, 35), thereby counteracting challenge 2 (*the organization objects to random assignment of persons or departments to the intervention or control group)*. Then, the intervention effect will be estimated using the propensity score through matching of individuals, stratification or regression adjustment (33, 36).

In a Finnish study, the propensity score was calculated for 24 000 persons in a cohort of public sector employees in municipalities and hospitals so they could be assigned to a multidisciplinary, vocational rehabilitation intervention to improve work ability (37). The propensity score was calculated using logistic regression analysis with 25 variables, including demographics (eg, gender), work characteristics (eg, work schedules), health risk indicators (eg, psychological distress), and health risk behaviors (eg, smoking status) (38). Once the propensity score was estimated, 859 employees who par-

ticipated in the intervention were matched by propensity score with 2426 controls, thereby excluding all other, unmatched employees in the entire cohort. The intervention showed adverse effects on perceived work ability and no beneficial effects on work disability: the risk of suboptimal work ability was somewhat higher after short- and long-term follow-up for participants than for controls (prevalence ratio 1.23 and 1.18, respectively) (37), while an earlier study showed that incident long-term work disability was about the same for participants and controls (hazard ratio 0.98) (38).

Some conditions need to be fulfilled before propensity scores can be considered as an alternative. The method assumes that all important prognostic variables are included and the model can be built perfectly (33, 34). If the propensity score is estimated or the covariates measured imperfectly, this bias may affect the estimated intervention effect (33). One way to cope with this problem is to construct different sets of propensity scores to test its robustness (39–41).

Instrumental variables method

The method of an instrumental variable is well known in the field of economics and applied to explore causal relationships between the intervention and an outcome in longitudinal studies (42). The method relies on finding a valid prediction variable, named "the instrument", that meets three assumptions: it (i) predicts the actual intervention received, (ii) is not directly related to the outcome, except by the direct effect of the intervention, and (iii) is not related to the outcome by any other measured or unmeasured path (42–44). Elovainio and colleagues recently investigated the association of job demands and job strain with perceived stress, psychological distress and sleeping problems among elderly care workers (45). Staffing level (ie, the ratio of the total number of nursing staff to the number of residents in the elderly care wards) appeared to be a strong instrument for both job demands and job strain, and instrumental regression analyses showed statistically significant associations with perceived stress and psychological distress. Self-reported job demands and job strain revealed the same results. An advantage of this method is that it provides a way to obtain a potentially unbiased estimate of treatment effect, even in the presence of strong unmeasured confounding (44). Since instrumental variables predict compliance to an intervention (or actual exposure) but have, by definition, no direct, independent effect on the outcome, the method of instrumental variables can reach the same effect as randomization (44) and thereby counteracts challenge 2 (*the organization objects to random assignment of persons or departments to the intervention or control group)*.

As an example of this method, Behncke (46) investigated the effects of retirement on various health outcomes (eg, self-assessed health, chronic condition, and biological measures). Of the 1439 individuals at baseline, 192 subjects retired during the two year follow-up. Behncke assumed that reaching the state pension age affected the retirement decision, but was not directly related to health outcomes. The analyses showed that state pension age was a good predictor for retirement and thus a strong instrument. The results of the instrumental analyses showed that retirement significantly increased the risk of being diagnosed with a chronic condition.

Choosing the correct instrument for the analysis is a crucial factor in this design. Weak instruments (ie, a low correlation between the instrument variable and intervention or exposure variable) lead to large standard errors resulting in imprecise and biased results when the sample size is small (47). Therefore, this method is particularly useful for large samples and in case of moderate confounding.

Multiple baseline design

In a multiple baseline design the same intervention is implemented at different time points across groups with pre- and post-measurements (48, 49). Outcome variables are measured in all groups at baseline. Then, one or more groups receive the intervention while others remain in the control condition. After sufficient time has passed for the intervention to affect the outcome, outcomes are again measured in all groups and the intervention is introduced in the next one or more groups (48–50). This procedure minimizes challenge 3 (*the organization wants to target all employees with an intervention)*. By sequentially introducing the intervention to groups, patterns of unexpected internal or external events can be studied; counteracting challenge 5 and 6 (*the organization is subject to internal/external change*). Compared to the RCT, fewer groups of participants are required in the multiple baseline design, since the group also acts as its own control (49); counteracting challenge 1 (*only few clusters exist to cluster the intervention at organizational level*). The design can be considered the non-randomized observational equivalent of the stepped-wedge design.

The evaluation of a behavioral contingency feedback intervention to increase attendance among 64 certified nursing assistants at three hospitals was conducted by applying a reversal (ie, ABA) multiple baseline design (51). The nine-week intervention was introduced across three groups at 16, 19, and 21 weeks after baseline measurement. All groups returned to the baseline situation (ie, A) after receiving the intervention (ie, B). The study ended with a final measurement after 39 weeks. The hospitals provided the research team with the working

schedules of the participants and their sickness absence records. The repeated measures analysis of variance showed that the total number of absent days per week decreased in the intervention period [mean 0.13, standard deviation (SD) 0.17] compared to baseline (mean 0.24, SD 0.19) and increased again after returning to the baseline situation (mean 0.24, SD 0.20).

The main statistical challenge in using the multiple baseline design is the high autocorrelation of repeated measurements over time, which can lead to imprecise estimates of the intervention effect (49). Autocorrelation can be removed by Auto-Regressive Integrated Moving Average or Independent Time Series Analysis modeling (49). Another challenge is achieving sufficient baseline stability, which includes enough data points for precise estimates (52). Third, the duration of the study should be sufficiently long to monitor external variations without interference of other influences, such as seasonal effects (49). Routinely collected data are an efficient means to establish a stable baseline over an extensive time period and this may even reduce data collection costs (49).

Interrupted time series design

In the interrupted time series design, a series of measurements is performed before and after implementation of the intervention at population level in order to detect whether the intervention has a significantly greater effect than the underlying secular trend, such as an economic, market or demographic trend (eg, the change in average body height of a population over time) (53). Whether the intervention had a significantly larger effect than any underlying trend is estimated by comparing the trend in the outcome after the intervention to the trend in the pre-intervention period (54, 55). Since randomization is not a prerequisite in this design, challenge 2 (*The organization objects to random assignment of persons or departments*) does not apply. The design is particularly relevant when using routinely collected data, such as workers' medical examinations, income insurance data, or workers' compensation data (26).

Farina and colleagues investigated the impact of national legislation on minimum safety and health requirements in 1999 on injuries at construction sites (56). Total and serious injury rates in the construction sector were calculated from 1994–2005, based on an integrated database (ie, Work History Italian Panel Salute). By applying segmented regression models that take into account secular trends and correct for any autocorrelation between the single observations, the results showed that the injury rates (per 10 000 weeks worked) decreased by 0.21 (95% CI -0.41– -0.01) per year more after the intervention than in the period before.

The main methodological concerns in applying the interrupted time series design for interventions are determining both the number of measurements before and after the intervention and the necessary time lags between measurements (eg, monthly or yearly data of sickness absence) to detect autocorrelations or secular trends (26, 57). Being able to determine specifically at what time point the intervention started is a precondition for applying the interrupted time series design (58).

Differences-in-differences

Differences-in-differences methods are common practice in economics to evaluate and interpret the effect of an inevitable change (eg, policy measure). In this design, observational data are used to compare the change in the outcome of a certain group that is subjected to an intervention at a specific time point to a change in the outcome in a group that is not exposed to this intervention (59). The method relies on finding a naturally occurring control group that mimics the properties of the intervention group and is therefore expected to follow the same time trend on the outcome as the intervention group would have in absence of the intervention (60). This design does not necessarily require measurements for the same individuals in each group over time, since repeated cross-sectional surveys can also be used (61). The intervention effect is calculated by subtracting the average change over time in the outcome variable in the control group from the average change in the intervention group. The design is thus an elegant way to study the internal or external changes that were named challenges earlier (*challenge 5 and 6*).

The differences-in-differences approach was applied to study the impact of a quality improvement intervention on reducing work disability, disability days, and disability and medical costs (62). The intervention firstly provided financial incentives to 512 health providers for faster adoption of occupational health best practices, and secondly focused on improvement of care coordination and disability management at patient level. A control group of 2297 providers with the same characteristics as the intervention group was constructed. Two cross sections of data were made, which included 33 910 workers' compensation claims in the baseline period (15 408 and 18 502 for the intervention and control groups, respectively) and 71 696 (31 520 and 40 176 in the intervention and control groups, respectively) claims during the follow-up period. Patients of the providers in the intervention group were significantly less likely to be off work after one year, leading to a reduction in disability days, and lower disability and medical costs.

As with the multiple baseline design and the interrupted time series design, the main methodological concern in this approach is the autocorrelation of the outcome (63). To deal with this issue, Bertrand and colleagues recommended conducting quite sophisti-

cated analyses, such as bootstrap techniques, when the number of groups is sufficiently large (63). Also, the differences-in-differences approach does not account for invariant factors and macro trends in one or both groups that might interfere with the outcome. Lastly, at the individual level, the impact of an intervention can be under- or overestimated due to unobserved, temporary and individual-specific events (60).

Regression discontinuity

The regression discontinuity design has been well established in economics over the last two decades, but not often applied in epidemiological studies. This design exploits a threshold or "cut-off" in a continuous variable used to assign treatment or intervention, and implies that individual whose assignment values lies "just above" or "just below" this threshold belong to the same population (64, 65) and thus can be compared to each other. The causal effects can be estimated by comparing the outcome between the two groups (66), assuming that subjects are not able to manipulate the threshold value. Hence, challenge 1, 2, and 3 concerning randomization and control group are minimized.

The causal effect of extending unemployment benefit duration on unemployment duration and post-unemployment outcomes was estimated in a regression discontinuity design (67). A sharp discontinuity for age could be used, since the maximum duration of unemployment benefits increases from 12 to 18 months at the age of 45. Age was considered the threshold value, ie, the assignment variable. The study population consisted of 3432 men (44–46 years) and 3784 women (43.5–46.5 years) who were unemployed in the period from 2001–2003. By including a dummy for being exposed (ie, being >45 years old), the exit rates from employment and unemployment in the group aged >45 years were compared to the exit rates from those in the control group. The hazard rates showed that a shorter duration of unemployment benefit was associated with a higher probability of entering paid employment.

The regression discontinuity design is only appropriate when treatment is applied to a strictly defined rule, linked to a continuously measured variable (such as duration of unemployment benefit in the example above) (66). The assumption that individuals around the threshold are similar is often debatable (64). Other important factors to consider when applying this design are the possibility of change over time in the assignment variable and the unequal distribution of missing data between the two groups. Applying this design requires larger sample sizes than an RCT to achieve sufficient statistical power (68). The feasibility of this design can be improved by using routine clinical or administrative data (66).

Discussion

This article demonstrated the appropriateness of research designs other than the RCT for the evaluation of occupational health interventions. Studies wherein these research designs have been applied successfully showed that the most fundamental research question in intervention research could be answered, ie, did change actually occur as a result of the intervention? The designs were either experimental in nature (ie, stepped wedge) or observational (ie, propensity scores, instrumental variables, multiple baseline design, interrupted time series, difference-in-difference, and regression discontinuity).

Some of the alternative designs (eg, multiple baseline design) require using more complex statistical models that may contain a relatively large number of parameters in order to account for heterogeneity across clusters. In these cases, larger sample sizes might be needed than would be the case for individually based RCT. Furthermore, in any intervention evaluation, it seems worthwhile to determine systematically how implementation influenced the results by conducting a process evaluation. A well-known implementation model for public health and community-based interventions is the RE-AIM framework, which assesses reach, efficacy, adoption, implementation, and maintenance (69). Nielsen and Randall's implementation model might be more helpful for organizational-level occupational health interventions since it additionally takes into account the mental models (ie, readiness for change and perception) of those involved (70).

Even though several researchers have acknowledged that conducting an RCT on a complex intervention within an occupational health context is not always preferable, the described alternative designs are not yet widely adopted in occupational health. This could be explained by unfamiliarity of researchers with the alternatives and their advantages and disadvantages compared to the RCT, or researchers feeling pressured to apply an RCT to maximize the possibility for publication. Hopefully, this article serves as a nudge for colleagues to consider alternative research designs for the evaluation of interventions. This article also aimed to provide the necessary information to decide on selecting the most appropriate design to answer the research question, with the highest level of internal and external validity possible and the lowest costs. Designs using observational data, for instance, are particularly useful for organizational interventions or policy measures with availability of sufficient administrative data allowing for a timely evaluation of the impact of such interventions. Observational designs may be especially applicable to research in dynamic work contexts characterized by eg, high turnover, organizational restructuring, or internal mobility. While the RCT is based

on a fixed cohort whereby individuals are enrolled at the same time (ie, the start of the study) and followed up for a similar period, this may be difficult when conducting an RCT in organizations with a high annual turnover of personnel. Some alternative designs are based on dynamic cohorts whereby individuals can enter and leave the cohort at different times, eg, the designs based on repeated cross-sectional data (see table 2). This may be an additional advantage to consider an observational design over an RCT.

The societal trend of big data deserves to be mentioned at this point. Some have proclaimed the current period, with its digitized patient records in large databases, to be an "open information era" as a result of public institution's and government's increased transparency (71). Research can benefit from the readily accessible data this "era" yields by combining large amounts of information gathered for different purposes via different devices or media (ie, big data, so called for its variety, volume, and velocity) (72). In doing so, we can discover correlations that would not be discovered in carefully constructed evaluations, which are typically set out to test causal relations. Big data are thus especially of interest for the described alternative research designs drawing on routinely collected data.

The research designs described in this article are appropriate to evaluate the effect of an intervention that is noticeable within a period of months to several years. However, the time lag between the intervention and the consequences for health can take many more years (eg, the effect of an intervention to reduce occupational exposure to dust on diseases such as silicosis and COPD). In these situations, other designs need to be considered, such as health impact assessment (HIA), which simulates the development of illness over time, based on the combined estimate of three models on: stage of the disease, the effect of exposure on stages of disease, and population characteristics. Meijster and colleagues combined a multi-stage model of respiratory problems, exposure to flour dust and allergens, and career length and influx of new workers, to estimate respiratory health outcomes of workers in the bakery sector (73). The probability on transitioning to the next stage of disease, per unit of exposure, per year was calculated, so that incidence could be determined. The combined model demonstrated how respiratory problems develop over time and how exposure and population characteristics contributed, eg, a mean latency period of 10.3 years (95% CI 8.3–12.3) for developing respiratory symptoms in bakers was predicted (73). Even though the RCT is still preferred as design for interventions targeted at individual level, this article provides an overview of appropriate alternatives when a group level intervention is applied, or if methodological or feasibility issues are encountered in an individual RCT that obscure the intervention-outcome relationship. The choice of the most appropriate design will be guided by the specific research question, complexity of the intervention, data available, context, and costs. Moreover, researchers conducting systematic reviews should not neglect evidence from studies applying alternative research designs. They should broaden their inclusion criteria towards observational studies with appropriate designs. When these alternative designs are applied more often, further research is necessary on the development and implementation of a guideline to improve the quality of reporting non-randomized controlled trials. We highly recommend to adopt and further explore the possibilities of both experimental alternatives and alternatives based on observational data for the evaluation of occupational health interventions.

Competing interests

The authors declare they have no competing interests.

Acknowledgment

The authors thank Noortje Wiezer for her valuable suggestions on earlier drafts of this narrative review.

References

1. Crofton JW, Clegg JW, Mitchison DA, Hurford JV, Douglas Smith BJ, Snell WE, et al. STREPTOMYCIN treatment of pulmonary tuberculosis. A Medical Research Council Investigation. Brit Med J. 1948;2(4582):769–82. http://dx.doi.org/10.1136/bmj.2.4582.769.

2. Martimo KP, Shiri R, Miranda H, Ketola R, Varonen H, Viikari-Juntura E. Effectiveness of an ergonomic intervention on the productivity of workers with upper-extremity disorders - A randomized controlled trial. Scand J Work Environ Health. 2010;36(1):25–33. http://dx.doi.org/10.5271/sjweh.2880.

3. Wolff N. Using randomized controlled trials to evaluate socially complex services: problems, challenges and recommendations. J Ment Health Policy Econ. 2000;3(2):97–109. http://dx.doi.org/10.1002/1099-176X(200006)3:2<97::AID-MHP77>3.0.CO;2-S.

4. Blackwood B. Methodological issues in evaluating complex healthcare interventions. J Adv Nurs. 2006;54(5):612–22. http://dx.doi.org/10.1111/j.1365-2648.2006.03869.x.

5. Hawe P, Shiell A, Riley T. Important considerations for standardizing complex interventions. J Adv Nurs. 2008;62(2):267. http://dx.doi.org/10.1111/j.1365-2648.2008.04686.x.

6. Glasgow RE, Klesges LM, Dzewaltowski DA, Bull SS,

Estabrooks P. The future of health behavior change research: what is needed to improve translation of research into health promotion practice? Ann Behav Med. 2004;27(1):3–12. http://dx.doi.org/10.1207/s15324796abm2701_2.

7. Campbell NC, Murray E, Darbyshire J, Emery J, Farmer A, Griffiths F, et al. Designing and evaluating complex interventions to improve health care. Brit Med J. 2007;334(7591):455–9. http://dx.doi.org/10.1136/bmj.39108.379965.BE.

8. Shepperd S, Lewin S, Straus S, Clarke M, Eccles MP, Fitzpatrick R, et al. Can we systematically review studies that evaluate complex interventions? PLoS Med. 2009;6(8):e1000086. http://dx.doi.org/10.1371/journal.pmed.1000086.

9. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: The new Medical Research Council guidance. Int J Nurs Stud. 2013;50(5):587–92. http://dx.doi.org/10.1016/j.ijnurstu.2012.09.010.

10. Glasgow RE, Lichtenstein E, Marcus AC. Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. Am J Public Health. 2003;93(8):1261–7. http://dx.doi.org/10.2105/AJPH.93.8.1261.

11. Ioannidis JPA, Haidich A-, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA-J Am Med Assoc. 2001;286(7):821–30. http://dx.doi.org/10.1001/jama.286.7.821.

12. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: Extension to cluster randomised trials. BMJ. 2012;345:e5661. http://dx.doi.org/10.1136/bmj.e5661.

13. Puffer S, Torgerson DJ, Watson J. Cluster randomized controlled trials. J Eval Clin Pract. 2005;11(5):479–83. http://dx.doi.org/10.1111/j.1365-2753.2005.00568.x.

14. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, et al. Improving the reporting of pragmatic trials: An extension of the CONSORT statement. Brit Med J. 2008;337(7680):1223–6. http://dx.doi.org/10.1136/bmj.a2390.

15. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. New Engl J Med. 2000;342(25):1887–92. http://dx.doi.org/10.1056/NEJM200006223422507.

16. Last JM, Spasoff RA, Harris SS, editors. A Dictionary of Epidemiology. 4th ed. New York: Oxford University Press; 2000.

17. Oude Hengel KM, Blatter BM, Joling CI, van der Beek AJ, Bongers PM. Effectiveness of an intervention at construction worksites on work engagement, social support, physical workload, and need for recovery: results from a cluster randomized controlled trial. BMC Public Health. 2012;12:1008. http://dx.doi.org/10.1186/1471-2458-12-1008.

18. Kristensen TS. Intervention studies in occupational epidemiology. Occup Environ Med. 2005;62(3):205–10. http://dx.doi.org/10.1136/oem.2004.016097.

19. Nielsen K, Taris TW, Cox T. The future of organizational interventions: Addressing the challenges of today's organizations. Work Stress. 2010;24(3):219–33. http://dx.doi.org/10.1080/02678373.2010.519176.

20. Bonell C, Fletcher A, Morton M, Lorenc T, Moore L. Realist randomised controlled trials: A new approach to evaluating complex public health interventions. Soc Sci Med. 2012;75(12):2299–306.

21. Griffiths A. Organizational interventions. Facing the limits of the natural science paradigm. Scand J Work Environ Health. 1999;25(6):589–96. http://dx.doi.org/10.5271/sjweh.485.

22. West SG, Duan N, Pequegnat W, Gaist P, Des Jarlais DC, Holtgrave D, et al. Alternatives to the randomized controlled trial. Am J Public Health. 2008;98(8):1359–66. http://dx.doi.org/10.2105/AJPH.2007.124446.

23. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007;28(2):182–91. http://dx.doi.org/10.1016/j.cct.2006.05.007.

24. Brown CA, Lilford RJ. The stepped wedge trial design: A systematic review. BMC Med Res Methodol. 2006;6:54. http://dx.doi.org/10.1186/1471-2288-6-54.

25. Mdege ND, Man M-, Taylor CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. J Clin Epidemiol. 2011;64(9):936–48. http://dx.doi.org/10.1016/j.jclinepi.2010.12.003.

26. Sanson-Fisher RW, D'Este CA, Carey ML, Noble N, Paul CL. Evaluation of systems-oriented public health interventions: Alternative research designs. Annu Rev Publ Health. 2014;35:9–27. http://dx.doi.org/10.1146/annurev-publhealth-032013-182445.

27. Zwijsen SA, Gerritsen DL, Eefsting JA, Smalbrugge M, Hertogh CM, Pot AM. Coming to grips with challenging behaviour: a cluster randomised controlled trial on the effects of a new care programme for challenging behaviour on burnout, job satisfaction and job demands of care staff on dementia special care units. Int J Nurs Stud. 2015;52(1):68–74. http://dx.doi.org/10.1016/j.ijnurstu.2014.10.003.

28. Kotz D, Spigt M, Arts ICW, Crutzen R, Viechtbauer W. Researchers should convince policy makers to perform a classic cluster randomized controlled trial instead of a stepped wedge design when an intervention is rolled out. J Clin Epidemiol. 2012;65(12):1255–6. http://dx.doi.org/10.1016/j.jclinepi.2012.06.016.

29. Kotz D, Spigt M, Arts ICW, Crutzen R, Viechtbauer W. Use of the stepped wedge design cannot be recommended: A critical appraisal and comparison with the classic cluster randomized controlled trial design. J Clin Epidemiol. 2012;65(12):1249–52. http://dx.doi.org/10.1016/j.jclinepi.2012.06.004.

30. D'Agostino RB,Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17(19):2265–81. http://dx.doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B.

31. Shadish WR, Luellen JK, Clark MH. Propensity Scores and Quasi-Experiments: A Testimony to the Practical Side of Lee Sechrest. In: Bootzin RR, McKnight PE, editors. Strengthening research methodology: Psychological measurement and evaluation. Washington, DC, US: American Psychological Association; 2006. p. 143–57. http://dx.doi.org/10.1037/11384-008.

32. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55. http://dx.doi.org/10.1093/biomet/70.1.41.

33. Luellen JK, Shadish WR, Clark MH. Propensity scores: an introduction and experimental test. Evaluation Rev. 2005;29(6):530–58. http://dx.doi.org/10.1177/0193841X05275596.

34. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivar Behav Res. 2011;46(3):399–424. http://dx.doi.org/10.1080/00273171.2011.568786.

35. Cousens S, Hargreaves J, Bonell C, Armstrong B, Thomas J, Kirkwood BR, et al. Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference. J Epidemiol Commun H. 2011;65(7):576–81. http://dx.doi.org/10.1136/jech.2008.082610.

36. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Methods. 2008;13(4):279–313. http://dx.doi.org/10.1037/a0014268.

37. Saltychev M, Laimi K, Oksanen T, Pentti J, Kivimaki M, Vahtera J. Does perceived work ability improve after a multidisciplinary preventive program in a population with no severe medical problems? The Finnish Public Sector Study. Scand J Work Environ Health. 2013;39(1):57–65. http://dx.doi.org/10.5271/sjweh.3298.

38. Saltychev M, Laimi K, El-Metwally A, Oksanen T, Pentti J, Virtanen M, et al. Effectiveness of multidisciplinary primary prevention in decreasing the risk of work disability in a low-risk population. Scand J Work Environ Health. 2012;38(1):27–37. http://dx.doi.org/10.5271/sjweh.3169.

39. Austin PC. Double propensity-score adjustment: A solution to design bias or bias due to incomplete matching. Stat Methods Med Res. 2014;1–22. http://dx.doi.org/10.1177/0962280214543508.

40. Steiner PM, Cook TD, Shadish WR. On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores. J Educ Behav Stat. 2011;36(2):213–36. http://dx.doi.org/10.3102/1076998610375835.

41. Steiner PM, Cook TD, Shadish WR, Clark MH. The importance of covariate selection in controlling for selection bias in observational studies. Psychol Methods. 2010;15(3):250–67. http://dx.doi.org/10.1037/a0018719.

42. Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. Annu Rev Publ Health. 1998;19:17–34. http://dx.doi.org/10.1146/annurev.publhealth.19.1.17.

43. Bennett DA. An introduction to instrumental variables analysis: part 1. Neuroepidemiology. 2010;35(3):237–40. http://dx.doi.org/10.1159/000319455.

44. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. J Clin Epidemiol. 2009;62(12):1226–32. http://dx.doi.org/10.1016/j.jclinepi.2008.12.005.

45. Elovainio M, Heponiemi T, Kuusio H, Jokela M, Aalto A, Pekkarinen L, et al. Job demands and job strain as risk factors for employee wellbeing in elderly care: an instrumental-variables analysis. Eur J Public Health. 2015;25(1):103–8. http://dx.doi.org/10.1093/eurpub/cku115.

46. Behncke S. Does retirement trigger ill health? Health Econ. 2012;21(3):282–300. http://dx.doi.org/10.1002/hec.1712.

47. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. Epidemiology. 2006;17(3):260–7. http://dx.doi.org/10.1097/01.ede.0000215160.88317.cb.

48. Moeyaert M, Ugille M, Ferron JM, Beretvas SN, Van den Noortgate W. Modeling external events in the three-level analysis of multiple-baseline across-participants designs: a simulation study. Behav Res Methods. 2013;45(2):547–59. http://dx.doi.org/10.3758/s13428-012-0274-1.

49. Hawkins NG, Sanson-Fisher RW, Shakeshaft A, D'Este C, Green LW. The multiple baseline design for evaluating population-based research. Am J Prev Med. 2007;33(2):162–8. http://dx.doi.org/10.1016/j.amepre.2007.03.020.

50. Rhoda DA, Murray DM, Andridge RR, Pennell ML, Hade EM. Studies with staggered starts: multiple baseline designs and group-randomized trials. Am J Public Health. 2011;101(11):2164–9. http://dx.doi.org/10.2105/AJPH.2011.300264.

51. Camden M, Ludwig T. Absenteeism in Health Care: Using Interlocking Behavioral Contingency Feedback to Increase Attendance With Certified Nursing Assistants. Journal of Organizational Behavior Management. 2013;33(3):165–84. http://dx.doi.org/10.1080/01608061.2013.814521.

52. Cuvo AJ. Multiple-baseline design in instructional research: pitfalls of measurement and procedural advantages. Am J Ment Def. 1979;84(3):219–28.

53. Biglan A, Ary D, Wagenaar AC. The value of interrupted time-series experiments for community intervention research. Prevention Science. 2000;1(1):31–49. http://dx.doi.org/10.1023/A:1010024016308.

54. Fretheim A, Soumerai SB, Zhang F, Oxman AD, Ross-Degnan D. Interrupted time-series analysis yielded an effect estimate concordant with the cluster-randomized controlled trial result. J Clin Epidemiol. 2013;66(8):883–7. http://dx.doi.org/10.1016/j.jclinepi.2013.03.016.

55. Zhang F, Wagner AK, Soumerai SB, Ross-Degnan D. Methods for estimating confidence intervals in interrupted time series analyses of health interventions. J Clin Epidemiol. 2009;62(2):143–8. http://dx.doi.org/10.1016/j.jclinepi.2008.08.007.

56. Farina E, Bena A, Pasqualini O, Costa G. Are regulations

effective in reducing construction injuries? An analysis of the Italian context. Occup Environ Med. 2013;70(9):611–6. http://dx.doi.org/10.1136/oemed-2012-101087.

57. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. Interrupted time series designs in health technology assessment: Lessons from two systematic reviews of behavior change strategies. Int J Technol Assess. 2003;19(4):613–23. http://dx.doi.org/10.1017/S0266462303000576.

58. Matowe LK, Leister CA, Crivera C, Korth-Bradley JM. Interrupted time series analysis in clinical research. Ann Pharmacother. 2003;37(7-8):1110–6. http://dx.doi.org/10.1345/aph.1A109.

59. Angrist J, Krueger A. Emperical strategies in Labor Economics. In: Ashenfelter O, Card D, editors. Handbook in Labor Economics. Elsevier; 2000. p. 1277-366.

60. Blundell R, Costa Dias M. Alternative Approaches to Evaluation in Empirical Microeconomics. J Hum Resour. 2009;44:565–640. http://dx.doi.org/10.1353/jhr.2009.0009.

61. Blundell R, MaCurdy T. Labor Supply. In: Ashenfelter O, Card D, editors. Handbook of Labor Economics. Elsevier; 2000. p. 1559–695.

62. Wickizer TM, Franklin G, Fulton-Kehoe D, Gluck J, Mootz R, Smith-Weller T, et al. Improving quality, preventing disability and reducing costs in workers' compensation healthcare: a population-based intervention study. Med Care. 2011;49(12):1105–11. http://dx.doi.org/10.1097/MLR.0b013e31823670e3.

63. Bertrand B, Duflo E, Mullainathan S. How much should we trust difference-in-difference estimates? The Quarterly Journal of Economics. 2004;119(1):249–75. http://dx.doi.org/10.1162/003355304772839588.

64. O'Keeffe AG, Geneletti S, Baio G, Sharples LD, Nazareth I, Petersen I. Regression discontinuity designs: an approach to the evaluation of treatment efficacy in primary care using observational data. Brit Med J. 2014;349:g5293. http://dx.doi.org/10.1136/bmj.g5293.

65. Moscoe E, Bor J, Bärnighausen T. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: A review of current and best practice. J Clin Epidemiol. 2015;68(2):122–33. http://dx.doi.org/10.1016/j.jclinepi.2014.06.021.

66. Bor J, Moscoe E, Mutevedzi P, Newell ML, Barnighausen T. Regression discontinuity designs in epidemiology: causal inference without randomized trials. Epidemiology. 2014;25(5):729–37. http://dx.doi.org/10.1097/EDE.0000000000000138.

67. Caliendo M, Tatsiramos K, Uhlendorff A. Benefit duration, unemployment duration and job match quality: a regression-discontinuity approach. J Appl Econom. 2013;28:604–27. http://dx.doi.org/10.1002/jae.2293.

68. Pennell ML, Hade EM, Murray DM, Rhoda DA. Cutoff designs for community-based intervention studies. Stat Med. 2011;30(15):1865–82. http://dx.doi.org/10.1002/sim.4237.

69. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health. 1999;89(9):1322–7. http://dx.doi.org/10.2105/AJPH.89.9.1322.

70. Nielsen K, Randall R. Opening the black box: Presenting a model for evaluating organizational-level interventions. European Journal of Work and Organizational Psychology. 2013;22(5):601–17. http://dx.doi.org/10.1080/1359432X.2012.690556.

71. Groves P, Kayyali B, Knott D, Van Kuiken S. The 'big data' revolution in healthcare: Accelarting value and innovation. Center for US Health System Reform Business Technology Office: McKinsey & Company; 2013.

72. Mooney SJ, Westreich DJ, El-Sayed AM. Commentary: epidemiology in the era of big data. Epidemiology. 2015;26(3):390–4. http://dx.doi.org/10.1097/EDE.0000000000000274.

73. Meijster T, Tielemans E, Heederik D. Effect of an intervention aimed at reducing the risk of allergic respiratory disease in bakers: Change in flour dust and fungal alpha-amylase levels. Occup Environ Med. 2009;66(8):543–9. http://dx.doi.org/10.1136/oem.2008.042564.

74. Smedley J, Harris EC, Cox V, Ntani G, Coggon D. Evaluation of a case management service to reduce sickness absence. Occup Med. 2013;63(2):89–95. http://dx.doi.org/10.1093/occmed/kqs223.

75. Kobayashi Y, Kaneyoshi A, Yokota A, Kawakami N. Effects of a worker participatory program for improving work environments on job stressors and mental health among workers: a controlled trial. J Occup Health. 2008;50(6):455–70. http://dx.doi.org/10.1539/joh.L7166.

76. Frykman M, Hasson H, Muntlin Athlin A, Von Thiele Schwarz U. Functions of behavior change interventions when implementing multi-professional teamwork at an emergency department: A comparative case study. BMC Health Services Research. 2014;14(1). http://dx.doi.org/10.1186/1472-6963-14-218.

77. Steenhuis I, van Assema P, Reubsaet A, Kok G. Process evaluation of two environmental nutrition programmes and an educational nutrition programme conducted at supermarkets and worksite cafeterias in the Netherlands. J Hum Nutr Diet. 2004;17(2):107–15. http://dx.doi.org/10.1111/j.1365-277X.2004.00507.x.

78. Oude Hengel KM, Blatter BM, Van Der Molen HF, Joling CI, Proper KI, Bongers PM, et al. Meeting the challenges of implementing an intervention to promote work ability and health-related quality of life at construction worksites: A process evaluation. J Occup Environ Med. 2011;53(12):1483–91. http://dx.doi.org/10.1097/JOM.0b013e3182398e03.

79. van Deurssen EH, Pronk A, Meijster T, Tielemans E, Heederik D, Oude Hengel KM. Process evaluation of an intervention program to reduce occupational quartz exposure among Dutch construction workers. J Occup Environ Med. 2015;57(4):428–35. http://dx.doi.org/10.1097/JOM.0000000000000382.

Received for publication: 24 February 2015