Effects of Change Interventions: What Kind of Evidence Do We Really Have?

Eric Barends, Barbara Janssen, Wouter ten Have and Steven ten Have

Journal of Applied Behavioral Science published online 28 January 2013
DOI: 10.1177/0021886312473152

The Journal of Applied Behavioral Science
XX(X) 1–23

DOI: 10.1177/0021886312473152
jabs.sagepub.com
SAGE

Effects of Change Interventions: What Kind of Evidence Do We Really Have?

Eric Barends[1], Barbara Janssen[1], Wouter ten Have[1], and Steven ten Have[1]

Abstract

Despite the popularity of organizational change management, the question arises whether its prescriptions are based on solid and convergent evidence. To answer this question, a systematic review was conducted of organizational change management research published in scholarly journals during the past 30 years. The databases ABI/INFORM, Business Source Premier, ERIC, and PsycINFO were searched for relevant studies. A total of 563 studies met the review's criteria. Assessment shows a predominance of one-shot studies with a low internal validity. Replication studies are rare. Findings suggest that scholars and practitioners should be sceptical regarding the body of research results in the field of organizational change management published to date. Prescriptions are offered for researchers, editors, and educators to develop a more solid body of evidence on organizational change management.

Keywords

organizational change, evidence, research design, replication, validity, systematic review

Introduction

Billions of dollars have been spent in the last two decades on management activities purportedly designed to change organizations. Virtually none of these efforts has any systematic monitoring or evaluation associated with it. This leads to an unfortunate state of affairs where the waxing and waning of organizational improvement

[1]VU University of Amsterdam, Amsterdam, Netherlands

Corresponding Author:
Eric Barends, VU University of Amsterdam, De Boelenlaan 1105, Room 2A17, 1081 HV Amsterdam, Netherlands
Email: e.barends@hotmail.com

> remedies are associated with limited understanding about what works and what does not and why.
>
> —Tichy (1983, p. 363)

Despite reportedly high failure rates (Beer & Nohria, 2000; Smith, 2002), the popularity of organizational change management (OCM) continues. In 1980, the database ABI/INFORM contained 426 articles on the subject in scholarly journals. This figure had risen to more than 1,700 by 1990 and to more than 6,800 by 2000. As of 2012, this database contained 20,000+ articles on OCM. Despite this publication boom, questions remain regarding whether and how well change management practices work. Thirty years ago, Noel Tichy (1983) was pessimistic on this subject. The issue remains whether this situation has improved. The present systematic review answers a question that must be resolved first. That is, whether change research is actually being conducted in a manner that can create the body of evidence necessary to provide conclusive findings regarding OCM's effectiveness.

The term *organizational change management* takes a variety of meanings. It came into use in the early 1970s and encompasses an array of concepts and methods that collectively address the question of how organizational change can be managed effectively (Beer & Nohria, 2000). Although many definitions have been presented (Bennis, Benne, & Chin, 1969; Burnes, 1992; Cummings & Worley, 1997; Schein, 1970), there is no generally accepted definition of OCM. Not surprisingly, the lack of a widely agreed-on definition and the enormous growth of new approaches in recent years blur the field's boundaries and make it difficult to describe (Cummings, 2004). The following definition draws on previous definitions in the field and guides the present systematic review:

> Organizational change management entails interventions intended to influence the task-related behaviour and associated results of an individual, team or entire organization.

This definition is intended to be broad reflecting the fact that management is an "integrative science," that is, a scientific endeavour composed of multiple disciplines (Noelle-Neumann & Schulz, 1971). The field of OCM incorporates research from numerous disciplines, including but not limited to economics, psychology, management science, and sociology. Thus, research in several disciplines must be considered in a systematic review of OCM.

Aim of This Review

The call for scientific substantiation of management theories is increasing markedly (Pfeffer & Sutton, 2006). This call seems particularly inspired by the question posed by Denise Rousseau in 2005, during her presidential address to the Academy of

Management: "Is there such thing as evidence-based management?" In the eponymous article published the following year, Rousseau (2006) called attention to a huge gap between science and practice, and that management decisions and interventions are thus often based on personal experience, intuition, or popular management models, rather than on the results of scientific research. For this reason, she proposed introducing to the field of management the principles of evidence-based practice already common in medicine, education, criminal justice, and social work. Since then, a large number of articles and books have been published on evidence-based management (Briner, Denyer, & Rousseau, 2009; Davenport & Marr, 2010; Latham, 2009; Lawler, 2007; Locke, 2009; Moss & Francis, 2007). A basic principle underlies all of evidence-based practice; that is, there must be available a solid and convergent body of reliable and valid evidence from multiple studies of the same constructs and interventions. From the perspective of OCM then the following subquestions must be addressed to answer the overarching question of this review:

1. *What research evaluating the effects of organizational change has been published in peer-reviewed scholarly journals in the past 30 years?*
2. *Which research designs are used?*
3. *What is the internal validity (i.e., control for bias and alternative explanations) of these research designs?*
4. *Which subject areas and variables are researched?*
5. *How are the outcomes of OCM measured?*
6. *To what extent are studies replicated to establish the validity and generalizability of their findings?*
7. *What do the answers to Questions 1 through 6 tell us about the body of evidence that exists on OCM?*

Although Tichy's (1983) observation was made over three decades ago, some contemporary scholars have raised similar concerns about the methodological quality of research in the field of OCM (Macy & Izumi, 1993; Pettigrew, Woodman, & Cameron, 2001; Woodman, 1989). Most of these concerns are based on a conventional review of the research literature. Such conventional reviews, however, are problematic. They are ad hoc, lacking both a systematic approach and clear criteria for inclusion. In addition, research results are not necessarily subjected to a critical appraisal to gauge the degree of confidence appropriate to them. Given the severe bias to which they are prone, conventional literature reviews are unsuitable for compiling an objective, comprehensive overview of the body of evidence on a topic (Antman, 1992; Bushman & Wells, 2001; Chalmers, Enkin, & Keirse, 1993; Fink, 1998). For this reason, we decided to conduct a systematic review to answer our research questions. The intention behind a systematic review is to identify as fully as possible all the scientific studies of relevance to a particular subject and to assess the validity and authority of the evidence of each study separately, based on such explicit criteria as research design, population, or outcome measures (Barends, ten Have, & Huisman, 2012). A well-specified approach

is applied to selecting studies and their methodological quality is assessed according to explicit criteria by independent raters (Higgins & Green, 2006; Petticrew & Roberts, 2006). A systematic review is therefore transparent, verifiable, and reproducible. The likelihood of bias is considerably less in a systematic review than in conventional literature reviews. Most systematic reviews focus on the findings of research, in order to guide practitioner decision making. In some cases, however, as in the present article, systematic reviews are limited to describing what kind of research has been done and thus comment on the relative strengths and weaknesses of how research on a specific topic has been constructed (Gough, 2012). Doing so can direct new empirical research to close gaps in a body of evidence, which is the aim of this review (Davies, 2004).

The Importance of Research Design

A study's design is central to the validity of its findings. The classification system for research designs most widely used is that of Campbell and his colleagues (Campbell & Stanley, 1966; Cook & Campbell, 1979). It classifies designs based on four kinds of validity: internal, statistical, construct, and external validity. When critically appraising research designs, all these types of validity are taken into account. However, to assess the methodological adequacy of OCM studies, internal validity is the primary concern. It indicates the extent of potential bias in a study's results and thus is a comment on the possibility of alternative explanations for them. Internal validity is an indicator of the extent that a cause-and-effect relationship between an intervention and its outcome is well-founded (Hoyle, Harris, & Judd, 2001). In effect, "Cause and effect can be established only through the proper research design: no amount of statistical hand waving can turn correlations into conclusions about causation" (Norman & Streiner, 2003). To determine which designs are the most robust in terms of internal validity, the so-called "levels of evidence" are used (Guyatt et al., 1995; Phillips et al., 2001). In the Campbellian classification, these levels describe the hierarchical ordering of research designs in terms of their internal validity (Figure 1). A study has a high internal validity when it fulfils the three conditions required for causal inference: covariation, time–order relationship, and elimination of plausible alternative causes (Shaugnessy, Zechmeister, & Zechmeister, 2006). The pure experiment or randomized controlled study is considered the design with the highest internal validity. Nonrandomized studies or quasi-experimental, observational, or correlation studies are regarded as research designs with lower internal validity (Campbell & Stanley, 1966; Cochran, 1965; Rosenbaum, 1995; Shadish, Cook, & Campbell, 2002). Examples of this type of research design include panel, cohort, and case control studies. Cross-sectional and case studies lack control over alternative explanations for their findings and are thus lower in the hierarchy of internal validity. An extensive overview of the application of research designs within the field of OCM is provided by Woodman, Bingham, and Yuan (2008).

It should be noted that the levels of internal validity as presented in Figure 1 are only relevant in assessing the validity of a cause-and-effect relationship that might

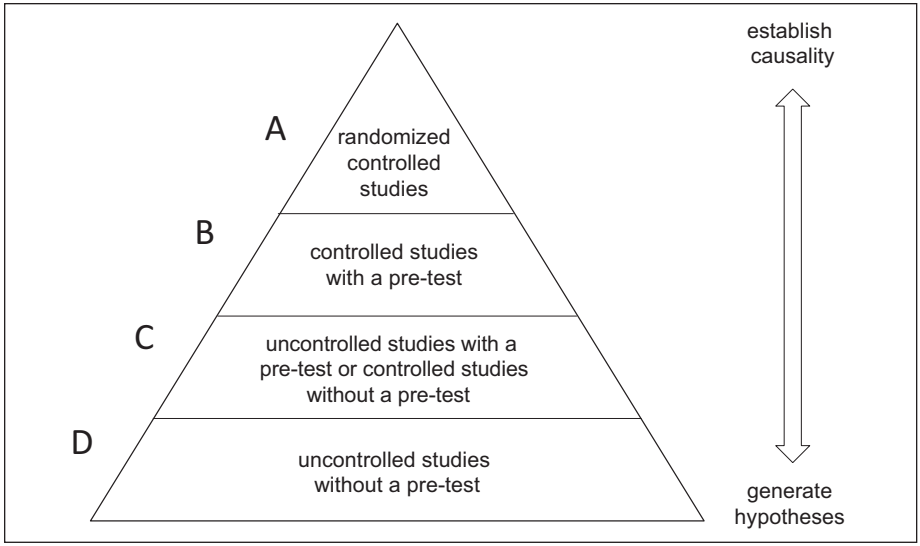

Figure 1. Levels of internal validity

exist between an intervention and its outcomes, which is the purpose of this review. Other research designs are more useful in other respects. Different types of research questions require different research designs (Petticrew & Roberts, 2003; Woodman, 1989). A case study for instance is a strong design for assessing why an effect has occurred or how an intervention might be (un)suitable in a particular context; it does a poor job of assessing the existence or strength of a cause-and-effect relationship (Trochim & Donnelly, 2007).

Body of Evidence

To answer a research question well, we need not a single study but a body of evidence. *All* research designs are flawed—though each is flawed differently (Creswell, Goodchild, & Turner, 1996; McGrath, 1981) For instance, research designs with a high internal validity, such as controlled studies, may be less subject to generalization, which restricts their practical usability. Cross-sectional surveys and case studies despite their lower internal validity can sometimes be more useful for identifying factors relevant to management practice. There is always a trade off between internal validity (precision in control and measurement) and external validity (generalizability with respect to populations and context). Researchers face a dilemma: maximize precision by reducing scope, so more is learned about less; or, maximize scope by accepting less precision so less is learned about more. McGrath made it clear more than 30 years ago that evidence accrual requires convergence of findings derived from multiple studies that investigate the same constructs and variables with different research

designs and in different contexts (McGrath, Martin, & Kulka, 1981). Building on McGrath's work, Edmondson and McManus (2007) noted that a body of evidence's stage of development is a key factor for determining the appropriate research method. In general, as an area of interest becomes more extensively studied, important contributions tend to take the form of quantitative tests with a high internal validity to identify critical independent, dependent, and moderating variables pertaining to an identified causal relationship. Conversely, the less is known about a specific topic, the more open-ended the research questions, requiring exploratory qualitative research with a lower internal validity to further shape the understanding of the topic (Edmondson & McManus, 2007).

The strongest contribution to a body of evidence comes from replication. Replication of studies is widely acknowledged as the touchstone of the scientific method, or put differently, "the Supreme Court of the scientific system" (Collins, 1985). Campbell and Stanley (1966) stated that "The experiments we do today, if successful, will need replication and cross-validation at other times under other conditions before they can become an established part of science" (p. 3). Thus, determining the scope and precision of research findings, replication plays a crucial role in ensuring the integrity of a body of evidence. Lack of replication means there is little or no chance for what is supposed to be the self-correcting nature of science to work (Neuliep, 1991). However, as McGrath (1981) and Edmondson and McManus (2007) pointed out, replicating studies in a different context or with a different research design is often not feasible for the individual researcher, who often has a strong preference and comfort level with certain methods. This means that to build a strong, convergent body of evidence, the focus should be on the collective research undertaken by many researchers instead of by a single individual or research team.

Method

To answer the seven subquestions described above, we conducted a systematic review following the specifications in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Moher, Liberati, Tetzlaff, & Altman, 2009).

Search Strategy

Four databases were used to identify relevant studies: ABI/INFORM Global from ProQuest, Business Source Premier from EBSCO, PsycINFO from Ovid, and the Education Resources Information Center (ERIC) database. These databases were selected because they contain publications dedicated to OCM as well as other journals from the field's supplier disciplines. A search was conducted using combinations of more than 50 different search terms compiled with the assistance of a business librarian.[1]

The following generic search filters were applied to all databases:

1. Scholarly journals, including peer reviewed
2. Published in the period 1980 to 2010
3. Articles in English

Selection of Studies

The following inclusion criteria were applied to the selection of studies:

1. *Organization*: Only studies related to an organization were included. An organization is defined as a formal collaboration between three or more persons who engage in activities to meet the demand for products and/or services in the organization's operating environment. Simulations of organizations (as used in experiments in artificial settings) were included.
2. *Interventions or moderators*: The only studies included were those involving an intervention intended to modify an organization's performance or moderating variables affecting the direction or strength of the outcome of an intervention.
3. *Dependent variable*: The only studies included were those involving interventions or moderators that directly or indirectly influenced the task-related responses and the associated performance of an individual, team, or entire organization.
4. *Measurement*: The only studies included were those in which the effect of interventions or moderators was measured or evaluated.
5. *Original studies*: Only original or primary studies were included. Reviews and meta-analyses were excluded.

Study selection took place in two phases. First, two reviewers independently screened the titles and abstracts of the 1895 studies identified for their relevance to this review. In case of doubt, lack of information, or disagreement, the study was included. Duplicate publications were removed. This first phase yielded 732 studies. Second, studies were selected for inclusion based on the full text of the article. Again, two reviewers worked independently. Study exclusions were tracked according to exclusion criterion and judgement. Where the two reviewers disagreed on inclusion, a third reviewer assessed the study's appropriateness for inclusion with no prior knowledge of the initial reviewers' assessments. The decision of the third reviewer was final. The second phase yielded 563 studies.[2]

Data Extraction and Classification

Data extraction was performed by two independent reviewers. The data extracted included research design, keywords, variables, and measurement scales. Each study's research design was categorized using Campbell's classification system (Petticrew & Roberts, 2006; Shadish et al., 2002). Any discrepancies were resolved

through discussion or by consulting a third party as needed. The following four levels of internal validity were used in the classification:

Level A: *Randomized controlled studies with a pretest*
- randomized controlled pretest–posttest designs

Level B: *Nonrandomized controlled studies with a pretest*
- nonrandomized controlled pretest–posttest designs
- cohort/panel studies
- case–control studies

Level C: *Controlled studies without a pretest or uncontrolled studies with a pretest*
- controlled posttest only designs
- uncontrolled pretest–posttest designs
- time series design

Level D: *Uncontrolled studies without pretest*
- uncontrolled posttest design with multiple posttests
- cross-sectional studies
- case studies

Studies were classified as controlled if one or more groups exposed to an intervention and/or moderating variable (experimental group) were compared with one or more groups not exposed or exposed to an alternative intervention and/or moderating variable (control group). Controlled studies were classified as randomized if the groups were chosen in a manner such that each participant (organization, team or employee) had equal chance of being assigned to either the experimental or the control group.

Controlled studies were classified as a pretest–posttest design where data were collected or specific characteristics measured both before and after exposure to the intervention. If there was only a posttest, then the study was classified as a posttest only design. The cohort or panel study classification applied to controlled studies involving groups (organizations, teams or employees) who were monitored for a long period (prospectively) to see if any difference arose between the groups (Rothman, Greenland, & Lash, 2008). The case–control study classification applied to controlled studies involving groups with a particular outcome (retrospective) that were compared with groups that did not experience this outcome (Schulz & Grimes, 2002).

Among the uncontrolled studies, a distinction was made between pretest–posttest designs, posttest-only designs with multiple posttests, cross-sectional studies, and case studies (Shadish et al., 2002). Uncontrolled studies in which the determinant and the outcome were measured simultaneously and uncontrolled studies with a single posttest were classified as cross-sectional. A study was classified as a case study where a large number (qualitative or quantitative) of aspects of a single organization or team were investigated in depth over a long period, without the use of a control group or pretest.

Keywords, Variables and Measurement Scales

To determine the most frequently researched subject areas, keywords for each individual study were retrieved and exported to Endnote. Keywords with little substantive meaning (such as "organization" or "study") were deleted. Excel 2011 and Word Counter 2.10 were used to calculate the frequency and cooccurrence with which keywords occurred across the 536 studies. Multiple occurrences of a keyword within the same study were treated as a single occurrence. In addition to the keywords, an analysis was also conducted of the variables and measurement scales that were employed by studies using a controlled or longitudinal design (Levels A, B, and C). For practical reasons, we excluded Level D studies in our analysis of variables: cross-sectional designs were excluded because they tend to include a wide range of variables, making coding difficult, and case study designs were excluded as in general these studies lack predefined variables. To minimize bias and eliminate ambiguity, two reviewers extracted the data from the studies independently. The variables identified were divided into four categories and classified into subcategories or subject areas based on their similarity in meaning. The individual results of the two reviewers were then combined and synthesized. Disagreements were resolved by discussion with a third reviewer.

Results

Research Design

Of the 563 studies included, a total of 75 used a control group (13%). The remaining 488 studies used no control group (87%). The results are summarized in Figure 2.

Of the 75 controlled studies, 11 used randomization. Although randomization is often considered incompatible with field research, the majority of randomized studies dealt with a field setting. An example of a randomized controlled (Level A) study in a field setting is the study by Shperling and Shirom (2005). Randomization was done at the organizational level (34 divisions were randomly assigned to an experimental or a control group), whereas statistical analysis was conducted at the individual level (Shperling & Shirom, 2005). Within the group of nonrandomized studies, 23 had a pretest–posttest design, 17 had a cohort or panel design, 14 a case–control design, and 10 studies a posttest-only design. Of the 23 studies with a pretest–posttest design, 17 featured multiple measurement points and data collected over a relatively long period of time, ranging from 1 to 6 years. An example is a (Level B) study regarding the effect of participative change programs on employees' job satisfaction and organizational trust by Macy, Peterson, and Norton (1989). In this research, surveys were administered three times at 18-month intervals to a group of 225 employees at an engineering division of a major electric power-generating company. A group of 227 matched employees at a similar engineering division within the same company in another city was selected by independent assessors as a comparison group. In this way, the

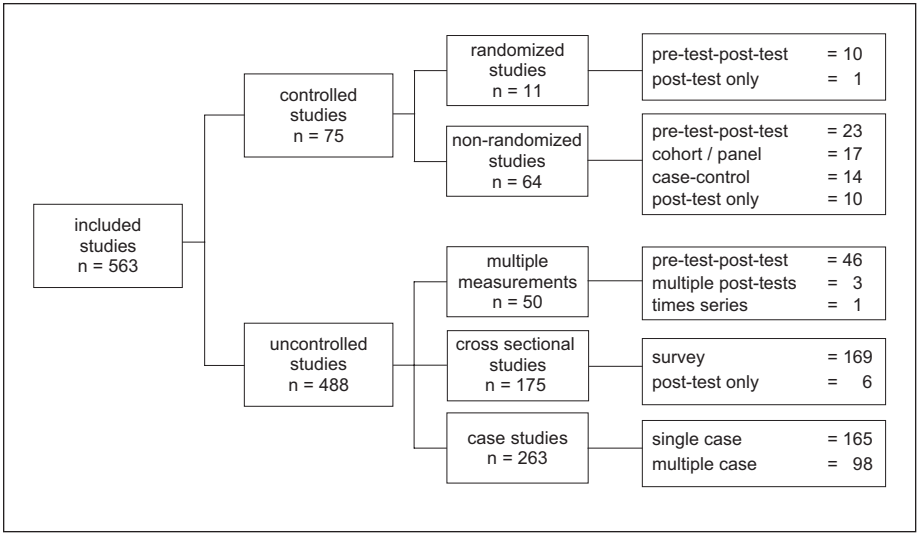

Figure 2. Number of studies and analysis of research designs

researchers attempted to compensate for the lack of randomization in order to establish causal relationships.

Within the 565 studies that were included, a total of 488 uncontrolled studies were identified. Of these 488 studies, 50 studies used more than one measurement point, including 46 studies with a pretest–posttest design, 3 studies with multiple posttests, and 1 study with a time series design. Many studies with a pretest–posttest design incorporated two or more groups (organizations, divisions, teams). However, these groups did not serve as a control, but were used instead to enhance external validity. An example is a (Level C) study of the effect of flexible work times on employee arrival and departure times by Ronen (Ronen, 1981). Its results suggest that employees, when given the opportunity to schedule their own workday, deviate only moderately from their preflextime arrival/departure times and that the number of late arrivals decreases considerably. Nonetheless, its lack of a control group means alternative explanations cannot be ruled out.

Among the 488 uncontrolled studies, a total of 175 cross-sectional studies were identified. The vast majority of these proved to be case studies. A total of 263 case studies were identified, including 165 single and 98 multiple case studies. Case studies were not only the dominant design among uncontrolled studies but also comprised more than 47% of the studies in this review. As noted above, case study designs are well-suited for theory generation or the identification of new phenomena, which suggests these studies are undertaken to "inspire other researchers to seek opportunities to expand their thinking and research" rather than to assess the effect of a change intervention (Lee, Mitschel, & Sablynski, 1999).

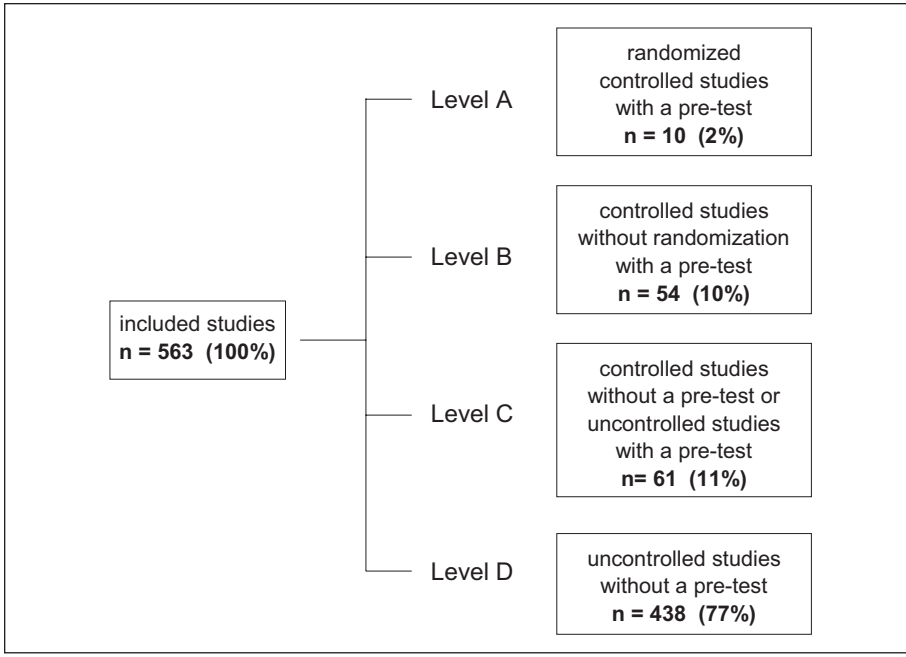

Figure 3. Number of studies and analysis of internal validity of evidence

Internal Validity

All studies were classified as Level A, B, C, or D based on their research design (Figure 3). Of the 563 studies included, only 10 qualified as Level A (2%) and 54 as Level B (10%). The remaining 499 studies (88%) had a weak internal validity.

To determine the trend, the ratio of controlled studies to uncontrolled studies was calculated for every 5 years between 1980 and 2009. As shown in Figure 4, the number of controlled studies over the past 30 years has remained relatively stable, but the number of uncontrolled studies has increased dramatically. This development means that the proportion of controlled studies dropped over the past 30 years from more than 30% in the early 1980s to just below 5% at the end of the first decade of the new millennium (Figure 4).

Keywords

A total of 1,171 unique keywords were identified in the 536 studies included.[3] The most frequently used keyword was Employee Attitudes (14%), followed by Corporate Culture (11%), Downsizing (9%), Leadership (8%), and Performance (7%). Employee Attitudes, Downsizing, Performance, and Employee Satisfaction are

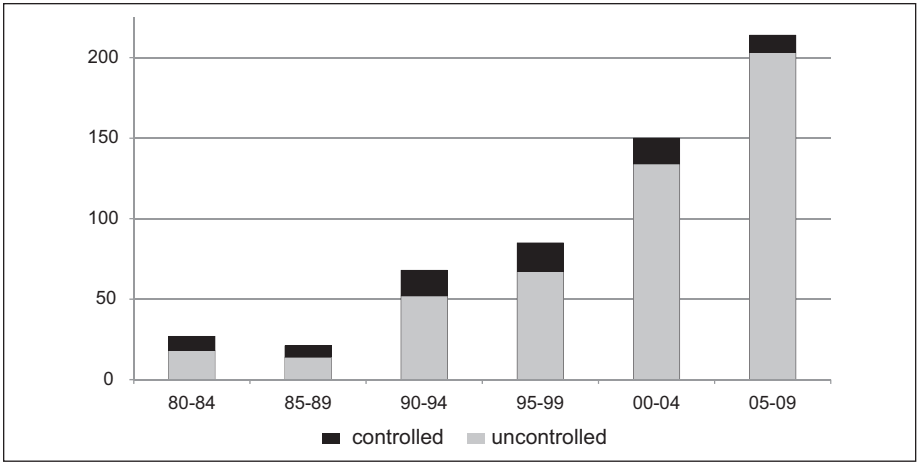

Figure 4. Controlled versus uncontrolled studies

used two to four times more often in studies with a controlled or longitudinal design. In comparison, the keywords Corporate Culture, Leadership Style, Strategic Management, and Knowledge Management occur two to four times as often in studies with a cross-sectional or case study design, suggesting that the research in these areas tends to be inductive or descriptive in nature. The number of studies sharing more than 2 similar keywords varies greatly, while the number sharing more than 3 similar keywords is minimal.

Variables

A total of 549 unique variables were identified in the 125 studies with a controlled or longitudinal design. Of these, 150 variables were classified as interventions, 124 variables as moderators, 246 variables as attitudinal outcomes, and 65 variables as objective performance measures.[4] The majority of the 125 studies measured attitudinal outcomes (76%), whereas 59 studies (47%) used objective performance measures. A total of 63 studies (50%) also measured the effect of moderating variables.

An overview of the most frequently researched interventions and moderators is given in Tables 1 and 2, respectively. Interventions in the subject area of "Downsizing" are the most frequent (25%), followed by interventions in the field of "Organizational Development" (17%) and "Performance Management" (15%). The subject areas of "Organizational Development," "Performance Management," "Strategic Change," and "Implementation of Change" cover a wide range of variables, suggesting that there is little to no replication of previous studies. In contrast, studies in the subject area of "Downsizing," "Change in top management," "Participation," "Job redesign," and "Mergers" cover a limited range of variables. However, the moderators paint a

Table 1. Analysis of Interventions of Controlled and/or Longitudinal Studies[a]

Subject area	Number of studies (n = 125)	Variables
Downsizing (reducing head count)	31 (25%)	Downsizing (19), restructuring (5), reorganization (6), major organizational change (3), drastic organizational change (1), radical organizational change (1), workforce reduction (1)
Human capacity building (skills, knowledge, and problem solving)	21 (17%)	Management training (3), problem solving (3), survey feedback (2), team building (2), team development (2), leadership training (2), empowerment (2), social reinforcement (1), intergroup development (1), 360 degrees feedback (1), individual coaching (1), people and career development (1), professional development (1), outdoor management education (1), learning organization intervention (1)
Performance & quality management (improving performance)	19 (15%)	Total quality management (7), incentives (4), goal setting (2), lean management (1), workflow formalization (1), process management (1), performance appraisal (1), reengineering (2), multifunctional teams (1), multitasking (1), quality improvements (1), quality circles (1)
Job redesign (enriching jobs)	10 (8%)	Work redesign (3), job enrichment (3), job redesign (2), work reorganisation (1), work restructuring (1), job enlargement (1)
Participation (increasing worker voice)	9 (7%)	Participation (4), participative decision making (2), participative management (1), power sharing (1), participative climate (1), participative versus authoritative leadership (1)
Change in top management (replacing senior leadership)	8 (6%)	Change in top management (3), CEO change (1), CEO succession (1), managerial succession (1), management replacement (1), executive migration /replacement (1), leader change (1)
Strategic change (changing mission or organizational activities)	6 (5%)	Strategic change (2), adoption to environmental change (1), change in strategy (1), repositioning (1), diversification (1), change in activities (1)
Implementation of change (changing operational practices)	6 (5%)	Implementation of change (1), implementation of new technology (1), implementation of change program (1), IT implementation (1), adoption of new technologies (1), implementation of organizational innovation (1)
Mergers and acquisitions (combining companies)	5 (4%)	Merger (3), acquisition (2), hostile takeover (1)
Other	14 (11%)	Other (13), such as relocation, technological change, privatization, CEO compensation, etc.

a. Counts are not mutually exclusive since a study may have more than one variable.

different picture: all four subcategories (employee characteristics, job characteristics, organization characteristics, and change process characteristics) show a wide variety of variables and none stand out as a major focus.

Table 2. Analysis of Moderating Variables in Controlled and/or Longitudinal Studies[a]

Subcategory	Number of studies (n = 125)	Variables
Employee characteristics	20 (16%)	Age (10), gender (3), education level (2), group characteristics (2), expectations toward the outcome (2), other (17), such as commitment, country of origin, involvement, motivation, preferences, self-efficacy, personality attributes, career stage, psychological flexibility, etc.
Job characteristics	16 (13%)	Job control (3), tenure (3), workload (3), job level (2), other (13), such as job autonomy, job demands, organizational level, physical demands, rank, promotional opportunities, routinization, developmental opportunities, transferable skills, external opportunities, etc.
Organization characteristics	19 (15%)	Firm age (4), organizational climate (3), other (15), such as industry experience, life cycle state, fit, organizational structure, managerial competence, trust in colleagues, freedom to suggest changes, rule enforcement, resource adequacy, shared vision, organizational culture, etc.
Change process characteristics	32 (26%)	Participation/participation in decision making (9), job insecurity (5), communication (4), perceived fairness (3), supervisory support (3), clarity of decision making (2), leadership (2), procedural justice (2), social support (2), understanding of the purpose (2), other (10), such as justification, decision-making influence, intensity of change intervention, met expectations, etc.

a. Counts are not mutually exclusive since a study may have more than one moderating variable.

Measurement Scales

The overview of the attitudinal outcomes in Table 3 shows that variables in the area of "Satisfaction" are the most frequently measured attitudinal outcomes (34%), followed by variables in the area of "Commitment" (20%), "Emotional Response" (20%), and "Well-being" (17%). However, the number of unique measurement scales is remarkably high. With the exception of the subcategories "Satisfaction" and "Commitment," the number of unique measurement scales is nearly equal to the number of studies, making it difficult to compare studies or to combine their outcomes to better estimate their effect size. The performance measures are presented in Table 4.

Table 3. Analysis of Attitudinal Outcomes of Controlled and/or Longitudinal Studies

Subcategory	Number of studies (n = 125)	Variables[a]	Number of unique measurement scales
Satisfaction	43 (34%)	Employee satisfaction (5), work satisfaction (3), job satisfaction (24), leader satisfaction (1), quality of work life (2), job motivation (1), work motivation (2), intrinsic reward satisfaction (1), morale (5), job attractiveness (1), work orientation (1), satisfaction with top management (1)	32
Commitment	25 (20%)	Organizational commitment (12), personal commitment (1), loyalty (1), organizational identification (1), organizational involvement (3), pride in organization (1), work centrality (1), intention to quit (1), intention to turnover (1), intention to remain (1), propensity to quit (1), turnover intention (1), outward orientation (1)	14
Emotional reactions	25 (20%)	Emotional reaction (3), response to workplace change (1), negative emotions (1), enthusiasm (1), social isolation (1), alienation (1), powerlessness (1), overconfidence (1), self-relevance (1), coping (3), perceived injustice (1), resistance to change (13), change acceptance, willingness to change, acceptance of change (5)	25
Well-being	21 (17%)	Employee well-being (1), occupational well-being (1), workplace well-being (1), general health (1), health symptoms (1), physical health (4), somatic complaints (1), psychological health (1), psychological distress (2), burnout (2), psychological withdrawal (1), psychology strain (1), job anxiety (4), mental health (2), depression (2), occupational well-being (1), stress (8), stress-related diseases (1)	20
Job characteristics	16 (13%)	Job autonomy (4), job control (3), job efficiency (1), workload (1), goal clarity (3), task uncertainty (1), role clarity (4), role overload (2), employee obligations, feedback (2)	15
Climate	16 (13%)	Organizational climate (5), conflict level (2), supportive leadership (2), group climate (1), social climate (1), open discussion (1), peer relationships (1), professional interaction (1), respect for authority of leaders (1), relationship with supervisor (1), group cohesiveness (1), competitiveness (1), organizational morale (1)	15
Trust	10 (8%)	Trust (4), organizational trust (2), management trust (2), workgroup trust (1), interpersonal trust (1)	8
Other	29 (23%)	Other (29), such as perception of change outcome, organizational learning, vision awareness, customer orientation, communication, goal attainment, task perception, corporate culture, etc.	29

a. Counts are not mutually exclusive since a study may have more than one outcome measure.

Table 4. Analysis of Performance Measures of Controlled and/or Longitudinal Studies[a]

Subcategory	Number of studies (n = 125)	Performance measures
Individual or team performance	23 (18%)	Performance (13), productivity (4), innovativeness (2), errors (2), product quality (2), other (15), such as grades, budget, patient safety, customer satisfaction, etc.
Organizational performance	37 (30%)	Absenteeism (9), productivity (6), performance (4), turnover (3), sales (3), innovativeness (2), return on assets (2), other (26), such as costs, market value, level of diversification, organizational mortality, customer satisfaction, safety, etc.

a. Counts are not mutually exclusive since a study may have more than one performance measure.

Discussion

We draw three conclusions from this review's results. First, the number of OCM studies has dramatically grown in the past 30 years. At the same time, the field's methodological repertoire remains limited. For example, the case control design seems to be relatively unknown in OCM: a total of only 14 studies (2%) featured this design. The same applies to time series (1%) and cohort/panel study designs (3%). Relatively few controlled studies are conducted in the field: Only 13% used a control group. In sharp contrast is the number of cross-sectional and case studies: More than 77% of studies made use of these designs.

Second, the internal validity of studies in the field of OCM tends to be low: of the 563 studies included, only 10 studies qualified as Level A (2%) and only 54 as Level B (10%). The remaining 88% are studies with a moderate to weak internal validity. Even more disturbing is the fact that the relative proportion of controlled studies into the effectiveness of interventions and/ or moderators within the field has *decreased* dramatically over the past 30 years, from more than 30% in the early 1980s to just less than 5% in the past decade.

Third, studies of OCM are quite heterogeneous in terms of content. Downsizing, Performance Management, and Organizational Development are the most researched subject areas in the field of OCM. Satisfaction, Commitment, Well-being, and Emotional Response are the most frequently measured attitudinal outcomes and Absenteeism, Performance, and Productivity are the most frequently measured "hard" outcome variables. However, taken together, the limited number of studies with similar keywords, the wide range of variables within subcategories, and large number of unique scales used to measure the outcome all suggest there is little to no replication in OCM.

Implications

If one accepts McGrath's (1981) fundamental premise that the success of the research enterprise is to be judged in terms of how well researchers seek convergence of substance among complementary research methods, then the outcome of this systematic review leaves us with a sense of limited success regarding OCM's capacity to answer fundamental questions about what works (and what does not). This review indicates that the field's research is limited by the dominance of one-shot studies and seldom addresses the same intervention more than once. As this may be appropriate for a field where theory is nascent and topics have attracted little research, it is unacceptable for a mature field with a research tradition of 50+ years. The field's lack of replication is particularly disturbing. OCM stands in contrast to fields such as medicine where research is often repeated and under different conditions in order to obtain the highest level of both internal and external validity. Research activities in OCM seem to be isolated, unrelated, and fail to build on previous studies. Instead, the pursuit of novelty in research, the development of new conceptual frameworks, and the pursuit of newfangled constructs appear to drive the research agenda. As a result, we know increasingly less about more. This, of course, makes Noel Tichy's (1983) critical comment from 30 years ago more relevant than ever: "This leads to an unfortunate state of affairs where the waxing and waning of organizational improvement remedies are associated with limited understanding about what works and what does not and why" (p. 363). Given these outcome findings, practitioners should be sceptical about relying uncritically on research findings relevant to OCM as a basis for important decisions.

It must be noted that this review does not answer the question how this dismal state of affairs regarding the low quality of OCM's body of evidence came about. What larger forces are driving us to such poor scientific practice? The fully answer to this question is beyond the scope of this review. However, we will briefly address two explanations.

First, it has been argued that the dynamic nature of OCM makes it difficult if not impossible to use randomization (Bullock & Svyantek, 1987) or control groups. However, as this review clearly demonstrates, good examples of randomized and controlled designs, though scarce, can be found in OCM. Moreover, most of the barriers toward the use of randomization or control groups are not unique to OCM. In fact, research fields including medicine, economics, and psychology faces similar barriers. These fields use other research designs, such as cohort, case control, and time-series designs. Such research designs too can lead to robust empirical foundations, particularly when repeated frequently and under varying conditions (Petticrew & Roberts, 2003). However, as we demonstrated in our review, these designs are relatively underused in OCM.

Second, it is suggested that organizations are dynamic systems that do not lend themselves to "normal" science methods (Daft & Lewin, 1990; Dooley & Johnson, 1995). However, as Thomas Kuhn (1962) emphasized, "an extraordinary science must not simply be critical of the established normal science paradigm; it must also present an alternative".

Put differently, alternative research models that overcome the dual hurdles of internal and external validity must be available first before a postnormal science approach can be regarded as viable. Since our review clearly demonstrates that in OCM the "normal" science approach is not used to its full potential, extending its methodological repertoire with the above-mentioned controlled designs remains the best available option.

Limitations

The primary potential limitation of this review is selection bias. Relevant studies may have been missed because we included only those studies with the subject terms *organizational change* and/or *change management* in the databases we employed. There might be OCM studies, especially in the field of Organizational Development, that do not use that term (Abraham & Michie, 2008). Still, it is likely that the 563 studies we included constitute a representative sample of change management studies.

Another potential limitation is publication bias. Since the 1950s, scholars have noted that systematic reviews based only on published studies are subject to the "file drawer problem" (Dickersin, 1990; Egger & Smith, 1998; Rosenthal, 1979; Sterling, 1959). This problem refers to the underrepresentation of studies with negative or relatively small effects, potentially skewing the outcome of the systematic review. Since publication bias primarily affects the outcome (effect size) of studies available for inclusion, its influence on our findings are probably limited. The same applies to the decision to restrict this systematic review to English-language studies.

The most important limitation of this systematic review concerns biases due to misclassification. During data extraction and classification, it became apparent that the research methods employed are not always readily identifiable. A number of studies thus may have been erroneously classified. On the other hand, even in the unlikely event that 20% of the studies were incorrectly classified and ascribed a level of evidence that is too low, the percentage of studies in Levels A and B (controlled and with a pretest) would increase only from 12% to 14% and the percentage of C and D studies (uncontrolled and/or without pretest) would decline only from 88% to 86%.

How to Proceed?

The systematic and extensive search that is the basis of our review allows us to conclude with some authority that there is solid evidence for what many scholars and academics have long suspected: Despite its growing volume, OCM's body of evidence is of low quality and growing lower. This trend must be reversed. Drawing on the ecumenical attitude proposed by McGrath (1981) and Edmondson and McManus (2007), we urge OCM researchers to join forces to collectively create a convergent body of evidence that builds on previous studies. Doing so means replicating studies. It means following up previous research with new studies incorporating at least some previously used variables and measures in different research designs and contexts. Such replications can also extend past research by adding new variables. The current body of OCM evidence would benefit substantially by an increase in controlled stud-

ies of high internal validity that permit investigation of alternative explanations and refinement of previously observed effects.

There is strong evidence of bias against publishing replication studies in social research (Bornstein, 1990; Easley, Madden, & Dunn, 2000; Hubbard & Vetter, 1996; Neuliep & Crandall, 1990). When a researcher successfully replicates findings of previous studies, editors and reviewers can find it difficult to see this as an important contribution, particularly if it confirms previous findings (Bornstein, 1990). We call on journal editors in the field of OCM to seriously consider the recommendations suggested by Evanschitzky and Armstrong (2012):

1. Identify important studies in the field that should be replicated and invite researchers to do so.
2. Dedicate a separate section to replication research.
3. Guarantee well-conducted replication studies some form of publication, for instance, a short version in print and an extensive version online.
4. Appoint a replications editor, as has been done by the *Journal of Applied Econometrics*.

These recommendations will not be effective until researchers, editors, and reviewers come to appreciate the importance of replication for self-correction in science. Graduate students need to be taught early in their education that only the combined outcome of both original *and* replication studies provides the quality of evidence needed to support effective decision making in management practice, which is the shared goal of practitioners and academics.

Finally, we acknowledge that scholars and researchers should be offered more than an indictment of OCM research and a handful of recommendations. Our goal is to promote discussion within the field of OCM of how and why we got to this poor state of affairs and what should be done to overcome it. Accepting the current state of affairs is to guarantee that practitioners never conclusively know whether OCM practices work. In consequence, researchers in the field of OCM will never attain the scientific aspiration to both understand the world and be useful to it.

Acknowledgments

We would like to thank the three anonymous reviewers for their feedback and insightful comments. We would also like to thank Bart de Jong for his helpful comments on an early version of the manuscript. Finally, we are especially grateful to Denise Rousseau for her guidance, encouragement and support.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. An overview of the search terms used to search in ABI/INFORM and PsycINFO, not included because of space limitations, is available from the authors on request.
2. A graphical overview of the outcome of the selection process, not included because of the space limitations, *is available from the authors* on request.
3. A table with the 20 most frequently used keywords and how they are distributed by study design, not included because of space limitations, *is available from the authors* on request.
4. An overview of the type of variables employed by the 125 controlled and/or longitudinal studies, not included because of space limitations, *is available from the authors* on request.

References

Abraham, C., & Michie, S. (2008). A taxonomie of behavior change techniques used in interventions. *Health Psychology*, *27*, 379-387.

Antman, E. M. (1992). A comparison of results of meta-analyses of randomized controlled trials and reccommendations of clinical experts. *Journal of the American Medical Association*, *286*, 240-248.

Barends, E., ten Have, S., & Huisman, F. (2012). Learning from other evidence-based practices: The case of medicine. In D. M. Rousseau (Ed.), *The Oxford handbook of evidence-based management*. New York, NY: Oxford University Press.

Beer, M., & Nohria, N. (2000). Cracking the code of change. *Harvard Business Review*, *78*, 133-141.

Bennis, W., Benne, K., & Chin, R. (1969). *The planning of change* (2nd ed.). New York, NY: Holt, Rinehart & Winston.

Bornstein, R. (1990). Publication politics, Experimenter bias and the replicaton process in social science research. *Journal of Social Behaviour and Personality*, *5*(4), 71-81.

Briner, R., Denyer, D., & Rousseau, D. M. (2009). Evidence-based management: Concept cleanup time? *Academy of Management Perspectives*, *23*(4), 19-32.

Bullock, R. J., & Svyantek, D. J. (1987). The impossibility of using random strategies To study the organization development process. *Journal of Applied Behavioral Science*, *23*, 255-262.

Burnes, B. (1992). *Managing change: A strategic approach to organisational development and renewal*. London, England: Pitman.

Bushman, B., & Wells, G. (2001). Narrative impressions of literature: The availability bias and corrective properties of meta-analytic approaches. *Personality and Social Psycholgy Bulletin*, *27*, 1123-1130.

Campbell, D. T., & Stanley, J. L. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Chalmers, I., Enkin, M., & Keirse, M. (1993). Preparing and updating systematic reviews of randomized controlled trials of health care. *Millbank Quarterly*, *71*, 411-437.

Cochran, W. (1965). The planning of observational studies in human populations. *Journal of the Royal Statistic Society (Series A)*, *128*, 134-155.

Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. Beverley Hills, CA: Sage.

Cook, D. T., & Campbell, D. T. (1979). *Quasi-experimental design: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Creswell, J., Goodchild, L., & Turner, P. (1996). Integrated qualitative and quantitative research: Epistemology, history, and designs. In J. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 11, pp. 90-136). New York, NY: Agathon Press.

Cummings, T. (2004). Organization development and change. In J. Boonstra (Ed.), *Dynamics of organizational change and learning* (pp. 25-43). Chichester, England: Wiley.

Cummings, T. G., & Worley, C. G. (1997). *Organization development and change*. (6th ed.). Cincinnati OH: South-Western College.

Daft, R. L., & Lewin, A. Y. (1990). Can organization studies begin to break out of the normal science straitjacket? An editorial essay. *Organization Science*, *1*, 1-9.

Davenport, T., & Marr, B. (2010). *The intelligent company: Five steps to success with evidence-based management*. Chichester, England: Wiley.

Davies, P. (2004). Systematic reviews and the Campbell collaboration. In G. Thomas & R. Pring (Eds.), *Evidence-based practice in education* (pp. 21-33). Maidenhead, England: Open University Press.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *Journal of American Medical Association*, *263*, 1385-1389.

Dooley, K. J., & Johnson, L. J. (1995). TQM, chaos and complexity. *Human Systems Management*, *14*(4), 1-16.

Easley, R. W., Madden, C. S., & Dunn, M. G. (2000). Conducting marketing science. *Journal of Business Research*, *48*, 83-92.

Edmondson, A. C., & McManus, S. E. (2007). Methodological fit in management field research. *Academy of Management Review*, *32*, 1155-1179.

Egger, M., & Smith, G. (1998). Bias in location and selection of studies. *British Medical Journal*, *316*(7124), 61-66.

Evanschitzky, H., & Armstrong, S. J. (2012). Research with in-built replication: Comment and further suggestions for replication research. *Journal of Business Research*. Advance online publication. doi:10.1016/j.jbusres.2012.05.006

Fink, A. (1998). *Conducting research literature reviews: From paper to the Internet*. London, England: Sage.

Gough, D. (2012). *An introduction to systematic reviews*. London, England: Sage.

Guyatt, G., Sackett, D., Sinclair, J., Hayward, R., Cook, D., & Cook, R. (1995). User's guides to the medical literature. IX. A method for grading health care recommendations. *Journal of the American Medical Association*, *274*, 1800-1804.

Higgins, J., & Green, S. (Eds.). (2006). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0.). Chichester, England: Wiley.

Hoyle, R. H., Harris, M. J., & Judd, C. M. (2001). *Research methods in social relations* (7th ed.). Belmont, CA: Wadsworth.

Hubbard, R., & Vetter, D. E. (1996). An emperical comparison of published replication research in accounting, economics, finance, management and marketing. *Journal of Business Research*, *35*, 153-164.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Latham, G. (2009). *Becoming the evidence-based manager: How to put the science of management to work for you*. London, England: Davies-Black.

Lawler, E. I. (2007). Why HR practices are not evidence-based. *Academy of Management Journal*, *50*, 1033-1036.

Lee, T. W., Mitschel, T. R., & Sablynski, C. J. (1999). Qualitative research in organizational and vocational psychology. *Journal of Vocational Behavior*, *55*, 161-187.

Locke, E. (Ed.). (2009). *Handbook of principles of organizational behavior: Indispensable knowledge for evidence-based management*. Chichester, England: Wiley.

Macy, B. A., & Izumi, H. (1993). Organizational change, design, and work innovation: A meta-analysis of 131 North American field studies—1961-1991. In R. W. Woodman & W. A. Pasmore (Eds.), *Research in organizational change and development* (Vol. 7, pp. 235-313). Greenwich, CT: JAI Press.

Macy, B. A., Peterson, M. F., & Norton, L. W. (1989). A test of participation theory in a work re-design field setting: Degree of participation and comparison site contrasts. *Human Relations*, *42*, 1095-1165.

McGrath, J. (1981). Dilemmatics, the study of research choices and dilemmas. *American Behavioral Scientist*, *25*, 179-210.

McGrath, J., Martin, J., & Kulka, R. (1981). Some quasi-rules for making judgement calls in research. *American Behavioral Scientist*, *25*, 180-221.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*, *6*(7), e1000097.

Moss, S., & Francis, R. (2007). *The science of management: Fighting fads and fallacies with evidence-based practice*. Bowen Hills, Queensland, Australia: Australian Academic Press.

Neuliep, J. W. (1991). *Replication research in the social sciences*. London, England: Sage.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Personality and Social Psychology*, *5*(4), 85-90.

Noelle-Neumann, E., & Schulz, W. (1971). *Methoden der Publizistikwissenschaft* [Methods of mass communication]. Frankfurt am Main: Fischer Lexikon Publizistik.

Norman, G., & Streiner, D. (2003). *PDQ statistics*. Hamilton, Ontario, Canada: BC Decker.

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies and typologies: Horses for courses. *Journal of Epidemiology & Community Health*, *57*, 527-529.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford, England: Blackwell.

Pettigrew, A. M., Woodman, R. W., & Cameron, K. S. (2001). Studying organizational change and Development: Challenges for future research. *Academy of Management Journal*, *44*, 697-713.

Pfeffer, J., & Sutton, R. (2006). *Hard facts, Dangerous half-truths, and total nonsense*. Boston, MA: Harvard Business School Press.

Phillips, B., Ball, C., Sackett, D., Badenoch, D., Straus, S., Haynes, B., & Dawes, M. (2001). *Levels of evidence*. Retrieved from http://www.cebm.net/index.aspx?o=1025

Ronen, S. (1981). Arrival and departure patterns of public sector employees before and after implementation of flexitime. *Personnel Psychology*, *34*, 817-822.

Rosenbaum, P. (1995). *Observational studies*. New York, NY: Springer-Verlag.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Rothman, K., Greenland, S., & Lash, T. (2008). *Moderd epidemiology* (3rd ed.). Philidelphia, PA: Lippincott Williams & Wilkins.

Rousseau, D. M. (2006). Is there such a thing as evidence-based management? *Academy of Management Review*, *31*, 256-269.

Schein, E. H. (1970). *Organizational psychology* (2nd ed.). Engelwood Cliffs, NJ: Prentice-Hall.

Schulz, K. F., & Grimes, D. A. (2002). Case-control studies: research in reverse. *Lancet*, *359*(9304), 431-434.

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

Shaugnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2006). *Research methods in psychology* (7th ed.). New York, NY: McGraw-Hill.

Shperling, Z., & Shirom, A. (2005). A field experiment assessing the impact of the focused diagnosis intervention on job autonomy. *Journal of Applied Behavioral Science*, *41*, 222-240.

Smith, M. (2002). Success rates for different types of organizational change. *Performance Improvement*, *41*, 26-33.

Sterling, T. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance: Or vice versa. *Journal of the American Statistical Association*, *54*(285), 30-34.

Tichy, N. (1983). *Managing strategic change: Technical, political and cultural dynamics*. New York, NY: Wiley.

Trochim, W., & Donnelly, J. P. (2007). *The research methods knowledge base*. Mason, OH: Atomic Dog.

Woodman, R. W. (1989). Evaluation research on organizational change: Arguments for a combined paradigm approach. In R. W. Woodman & W. A. Pasmore (Eds.), *Research in organizational change and development* (Vol. 3, pp. 161-180). Greenwich, CT: JAI Press.

Woodman, R. W., Bingham, J. B., & Yuan, F. (2008). Assessing organization development interventions. In T. G. Cummings (Ed.), *Handbook of organization development* (pp. 187-207). Thousand Oaks, CA: Sage.

Bios

Eric Barends is a researcher at the VU University of Amsterdam and the director of the Center for Evidence-Based Management.

Barbara Janssen is a management consultant and a researcher at the VU University of Amsterdam.

Wouter ten Have is a management consultant and a teacher and researcher at the VU University of Amsterdam.

Steven ten Have is a management consultant and professor of Strategy and Change at the Strategy & Change at the VU University of Amsterdam.