Evaluating engineering design methods: taking inspiration from software engineering and the health sciences

Andreas Makoto Hein, Guillaume Lamé

INTERNATIONAL DESIGN CONFERENCE – DESIGN 2020
https://doi.org/10.1017/dsd.2020.317

EVALUATING ENGINEERING DESIGN METHODS: TAKING INSPIRATION FROM SOFTWARE ENGINEERING AND THE HEALTH SCIENCES

A. M. Hein ✉ and G. Lamé

CentraleSupélec, France

✉ andreas-makoto.hein@centralesupelec.fr

Abstract

Engineering design methods are typically evaluated via case studies, surveys, and experiments. Meanwhile, domains such as the health sciences as well as software engineering have developed further powerful evaluation approaches. The objective of this paper is to show how evaluation approaches from the health sciences and software engineering might further the evaluation of engineering design methods. We survey these approaches and show which approaches could be transferred to the evaluation of engineering design methods.

Keywords: design methods, design methodology, evaluation, empirical studies, research methodologies and methods

1. Introduction

Design methods are a core product of engineering design research (Gericke et al., 2017). However, many issues remain with the definition of design methods, and the evaluation of their impact (Gericke et al., 2017). The general aim of evaluation is "to determine merit, worth, value or significance" (Hawe et al., 2009). For design methods, this means: Does the method bring about the desired improvements in the design process? These improvements can relate to different aspects of the performance of the design method, such as the object of design (quality of the final design, cost of design, etc.) and the design process (development lead-time, capacity to find good solutions, etc.) (Badke-Schaub and Frankenberger, 1999; Dorst, 2008). This is particularly relevant for making rational decisions about which method to choose for solving a particular design problem (Fenton, 2001). However, even for widespread design methods such as Quality Function Deployment (QFD), empirical evidence for its effect in practice is mixed (Griffin, 1991). If even the most well-known and widespread engineering design methods lack clear evidence for their advertised impact in practice, this should bother us as a community.

A limited number of publications have dealt with the evaluation of design methods (Blessing and Chakrabarti, 2009; Frey and Dym, 2006; Seepersad et al., 2006), but the research community has not reached a consensus on how to proceed (Gericke et al., 2017). Possibly as a result, many papers that propose a method do not discuss its validity and impact (Barth et al., 2011; Blessing and Chakrabarti, 2009). This does not mean that design methods are in general not evaluated. Numerous publications deal with the empirical evaluation of specific design methods, mostly via case studies, surveys, and experimental setups (Bryant et al., 2006; Ferreira and Gil, 2012; Hein, 2016; Shi et al., 2019).

However, the main shortcoming is related to the evaluation of the effect of design methods in practice and how far these results can be generalized.
In this paper, we propose to look at evaluation practices in other fields; namely, software engineering and the health sciences. We review how evaluation is performed in these fields and explore what practices and approaches could be useful for evaluating engineering design methods. We select software engineering and health sciences, as they both deal with the evaluation of the effect of methods in often complex organizational settings.

2. Evaluating methods in software engineering

Software engineering methods are by their very nature closer to engineering design than the health sciences, as they focus on supporting the design of a technical artefact (software). Software engineering engages in various evaluation activities based on empirical data. Compared to engineering design, the domain has a longer and more extended tradition of empirical research. Evidence are the existence of a dedicated journal on this topic, the Empirical Software Engineering journal, which appears regularly since 1996. Furthermore, the number of papers on empirical software engineering (Keywords "software engineering", empirical - 364,000 results on google scholar) are much higher than for engineering design (Keywords: "engineering design", empirical – 126,000). Some of its proponents still criticize the maturity of their field and take inspiration from the medical sciences (Kitchenham et al., 2002; Pickard et al., 1998). Some of the evaluation approaches are (Easterbrook et al., 2008):

- Case studies (single, multiple) (Runeson et al., 2012)
- Surveys
- (Controlled) experiments
- Ethnographies
- Action research
- Meta-analyses (Pickard et al., 1998)

Most of these approaches are in widespread use in engineering design research for evaluating design methods such as case studies (Blessing and Chakrabarti, 2009) and surveys (Cristiano et al., 2001). Furthermore, ethnographies and action research are used for exploring the context within companies before, during, and after deployment of a design method (Bunning, 1995; Radcliffe and Harrison, 1994). However, not all approaches have already been used in engineering design research.

2.1. Meta-analyses in software engineering

An approach, which to our knowledge has found limited use in engineering design research are meta-analyses. Exceptions are Sio et al. (2015) and Cash (2018). Meta-analysis "is a technique for pooling data from different studies" (Hedges and Olkin, 2014; Pickard et al., 1998). The main objectives are to resolve uncertainty if studies disagree and to increase the confidence in the results of individual studies. Meta-analysis is considered powerful, as it allows for generalizing from individual studies. This is particularly relevant in software engineering, as results are difficult to generalize due to limitations in population selection (participants in studies and company context) and the difficulty to define constructs (precise specification of a software method to assure replicability) (Pickard et al., 1998). Furthermore, one key result we would like to get from a method is the size of its effect, called "effect size" in statistics. A precondition for meta-analysis is the use of a quantitative measure of effect size for each study.
For example, it is rather common that studies in software engineering have contradictory outcomes. For example, Dybå and Dingsøyr (2008) report inconclusive results from studies where the effect of using agile development methods is analyzed. The studies evaluated the effect on productivity and software quality. For both, the studies showed positive, indifferent, and negative results compared to an alternative method used as a control.
Several meta-analyses have been conducted in software engineering such as above-mentioned analysis on agile development methods, comprising 11 studies (Dybå and Dingsøyr, 2008), test-driven

development, comprising 27 studies (Rafique and Mišić, 2012), and pair programming, comprising 18 studies (Hannay et al., 2009). The fact that meta-analyses have been done in software engineering indicates that a sufficiently large number of statistical studies with regard to a specific method have been published.
The status quo in engineering design seems to be still far away from reaching this goal for several reasons. The large number of design methods is contrasted by the small number of statistical studies for one specific method, if there is a statistical study at all. Furthermore, measures for quantifying the effect of design methods are not well developed and no wider consensus exists on which measures to use.

2.2. Guidelines and roadmaps for empirical software engineering research

Several articles propose guidelines and roadmaps for empirical software engineering research, where method evaluation is an important element (Dybå et al., 2012; Easterbrook et al., 2008; Kitchenham et al., 2002; Ko et al., 2015; Perry et al., 2000; Runeson et al., 2012; Shull et al., 2007). Sjoberg et al. (2007) propose the following objectives for empirical software engineering research:

- More software engineering research should be based on the use of empirical methods;
- The quality, including relevance, of the studies using such methods should be increased;
- There should be more and better synthesis of empirical evidence; and more theories should be built and tested.

The issues mentioned in the last point have their analogue in engineering design research (Cash, 2018; Lamé, 2019). To achieve these objectives, Sjoberg et al. (2007) and Dingsøyr et al. (2008) propose:

- Increased competence regarding how to apply and combine alternative empirical methods;
- Tighter links between academia and industry;
- The development of common research agendas with a focus on empirical methods;
- More resources for empirical research;
- Providing more empirical research, primarily on experienced teams and organizations;
- Connecting better to existing streams of research in more established fields;
- Giving more attention to management-oriented approaches;
- Larger base of studies to provide opportunity for theory-building.

It seems that the software engineering community critically reviewed the state of practice of empirical software engineering in the early 2000s and developed roadmaps to improve the state of the art. A sub-community of empirical software engineering researchers exists with a dedicated journal and international conferences (*Empirical Software Engineering and Measurement*). The publication of several meta-analyses on specific software design methods (agile methods, pair programming, test-driven development) demonstrates a maturity which has not (yet) been achieved by the engineering design community.

3. Evaluating “complex interventions” in health sciences

Evaluation is a key research activity in medical and health sciences (Lamé, 2019). An important share of the research effort in these fields is spent on assessing the impact and effectiveness of proposed treatments and interventions on predefined outcomes. The objective is to establish a *causal relationship* between the use of a certain intervention in a given situation, and a change in a specified outcome.
Frey and Dym (2006) explored how approaches used to evaluate the effectiveness of medical treatments could be used to validate design methods. However, Frey and Dym’s review focuses narrowly on drugs. Therefore, their review does not cover other types of interventions that contribute to improve health, like computer decision support for health professionals, health promotion and population health interventions, or the introduction of new processes and organisations for delivering healthcare.
A broad range of interventions can be evaluated in health sciences, beyond pharmaceutical interventions. Two (related) dimensions can be used to describe this variety: (i) the extent to which the

intervention is affected by the context of implementation, and (ii) the number of components in the intervention.

As pointed out by (Frey and Dym, 2006), some interventions are very "bounded", in that they leave little room for variation due to personal interpretation or skill. Many injected drugs, for instance, are prescribed in defined doses and injected by professionals, so the intervention (the injection of a defined dose of a defined drug) is quite standard. However, other interventions allow more variation. For instance, surgical procedures need to be adapted to patients, and are also affected by surgeon's aptitudes. Similarly, training programmes can never be scripted in full detail (even if the message is fully specified, the tone and attitude of the person delivering it can only be prescribed) and will often be slightly adapted to local contexts.

Some interventions have only one component. This is the case when the comparison is between two injected drugs: the intervention is the replacement of one chemical agent by another. More complex interventions can have multiple parts, carried out by different people. A new pathway for a given disease is likely to cut across community care, primary care and secondary care, the organisation and availability of which can vary between locations. Implementing the pathway may also require training people and running a communication campaign, both of which are part of the intervention.

Interventions that are strongly affected by the context of implementation, and that comprise multiple components, are often labelled "complex interventions". For example, training programmes, health promotion programmes like smoking cessation programmes, or new pathways that organise the delivery of care between various providers combine multiple activities and require adaptation to the context of delivery, interpretation of the objectives and contents of the programme by those who will implement it, and coordination between multiple activities and providers. This type of intervention can be conceptualised as an event within a system, rather than as a package of discrete activities implemented in a static milieu (Hawe et al., 2009). The outcome of the intervention depends on how the system reacts to the event. This reaction is emergent and cannot be entirely prescribed or anticipated.

Design methods are arguably closer to complex health interventions than to simpler interventions such as drugs. Even the simplest and most standardised design methods need to be learnt, interpreted and adopted by users, often in an organisational context that requires collective adoption and organisational change to ensure implementation.

3.1. Traditional experimental and semi-experimental approaches to evaluation

Different methods exist to assess the impact of an intervention. A common notion in clinical research is that these methods can be ranked on a "hierarchy of evidence" (Merlin et al., 2009; Murad et al., 2016). This hierarchy is used as a heuristic to assess how strong the results of a study are based on the methods used to obtain them.

At the top of this hierarchy are controlled experiments, often called "controlled trials" in medicine. In these studies, study participants are randomly assigned to one of two groups. One of the groups is exposed to the intervention, while the other group (termed the control group) is not. Outcomes are measured before and after the intervention. Ideally, those involved in the process (patients and medical staff) do not know in which group they are. Because of the random allocation, groups can be equivalent, and blinding participants to what they receive reduces psychological effects. This study design emulates laboratory experiments in the physical and natural sciences. The researcher has strong control on what happens, the control and intervention groups are similar.

Below randomised controlled trials are pseudo-randomised controlled trials, where allocation between the groups is pseudo-random. For instance, a patient could be allocated to the intervention or the control group based on the day of the week when they agree to participate in the study.

Then come comparative studies with concurrent control groups, but without randomisation. These include controlled before-and-after studies, where allocation to the groups is not controlled by the researcher. The groups can therefore not be assumed to be equivalent. Outcomes are still measured before and after the intervention, and they are compared between the control and the intervention group.

Finally, interrupted time-series without concurrent control collect multiple data points for a single group that receives the intervention. The effect of the intervention is assessed by comparing what happens after the intervention to the trend before the intervention.

The choice of one of these designs does not prejudge of the setting in which the study will be conducted. Randomised controlled trials can be to assess the effect of treatments in the open environment of a hospital, but they can also be used to assess the effect of training interventions in simulated environments (Lamé and Dixon-Woods, in press), a controlled setting that is closer to the laboratory settings in physical sciences.
A major issue is to consolidate learning from different programmes. Similar interventions are tested across the globe and these evaluations of the same intervention can give contradicting results. To overcome this issue and synthesise results, health sciences have developed standardised procedures for systematically reviewing the literature and synthesising evidence, including through quantitative meta-analysis (Lamé, 2019).

3.2. Challenges of evaluating complex interventions with traditional methods

In the traditional use of the experimental and quasi-experimental designs described above, the intervention is treated as a black box. Outcomes are measured before and after, and changes can be attributed to exposure to the intervention, but when sticking to these designs without further investigation there is no way to explain how the intervention generates its effects. This is particularly important in the case of complex interventions that comprise multiple components and are strongly dependent on context.
When interventions have multiple components, black box evaluation does not allow to distinguish the impact of each of these components, or their interactions. A black box evaluation will only show the aggregated effect of the system of components. However, it may be that a subset of components is entirely responsible for the results, while the rest of the components have no impact. It can also be that some components are critical and require specific attention.
Context-dependency makes it difficult to understand if the outcomes of the evaluation are due to the intrinsic design of the intervention, or to the way it was implemented in a specific context. For example, in the case of a smoking prevention programme in schools, part of the intervention can be standardised package (visuals, leaflets, agenda of the sessions), but the setting of the intervention (location, time) and the way it is presented to participants by school staff will vary between schools, and the way sessions will happen will be the result of the interaction between the intervenors and the pupils in the room and will be different each time. Therefore, if an evaluation of this smoking cessation programme shows no effect (the number of students who smoke or start smoking remains stable), it may be because the intervention itself is ineffective (what was designed does not work in practice), but it may also be because the intervention was not implemented appropriately (what happened was not what was designed) (Craig et al., 2008). It may also be that certain contextual factors inhibit or counteract the effects of the intervention.

3.3. Methods for evaluating complex interventions

In the case of complex interventions, the question evaluators need to answer is not only, "does this intervention work?", but "what works, for whom, under which circumstances?". This requires moving beyond black box evaluation to gain an understanding of what happens "inside" the intervention.
Although Frey and Dym (2006) argue that theory does not play a great role in the evaluation of pharmaceutical interventions, theory-driven evaluation is increasingly used for complex interventions. In theory-driven evaluation, evaluators build a theoretical model of how the combination of all components of the intervention is thought to contribute to generating certain outcomes (Breuer et al., 2016; De Silva et al., 2014). Evaluators can then assess each of these causal links during the evaluation. Such theories are sometimes called "programme theories".
Theories in this context do not need to be grand abstract constructions. Building on existing research evidence and on the intervention designers' rationale for creating the intervention, they show how different aspects of the intervention interact and how they interact with the context of implementation to generate outcomes. They can often be represented by logic models or causal maps. The objective of the evaluation is to test the underlying theoretical model, rather than the intervention itself, which will vary from one context of implementation to the next.

As part of theory-driven evaluation, process evaluation (Moore et al., 2015) is the monitoring of intervention delivery to scrutinise how the intervention was implemented (did things work as planned?), to analyse what mechanisms link components of the intervention to outcomes, and to understand how the local context affected intervention delivery and outcomes. 80% of evaluations of complex interventions include some form of process evaluation (Minary et al., 2019).
To capture contextual influences and participants' reactions to the intervention, process evaluations can combine quantitative and qualitative methods. In particular, qualitative methods can be used to elicit participants' lived experience and different perspectives on the intervention (Green and Britten, 1998).

3.4. Lessons for engineering design

Frey and Dym (2006) have already argued that engineering design could learn from experimental approaches used in medicine. Our updated review shades new light and brings more insights on what could be transferred from health sciences.
Compared to the types of interventions that exist in health sciences, engineering design methods err on the side of complex interventions. Even when they are simple and well packaged, methods need to be learnt, interpreted and adopted by users, often in an organisational context that requires a collective adoption. Therefore, we should look for evaluation methods that are used for evaluating complex health interventions.
In particular, the theory-driven approach to evaluation has been developed to overcome the limitations of black-box evaluation. It requires careful delineation of why and how the intervention is hoped to generate desired outcomes. Existing tools in design research could already support a theory-driven approach to evaluation. For example, Design Research Methodology (DRM) (Blessing and Chakrabarti, 2009) proposes different types of graphic models to elicit and represent causal chains which design methods seek to affect. However, DRM diagrams focus on the causal chain with which the design method interferes, but they do not describe causal chains inside the method. They also do not account for implementation strategies.
To apply a theory-driven approach to evaluate design methods, methods and their implementation strategy need to be considered as one intervention system, the components of this system described, and their relationships specified. This supposes that the evaluation bears not only on the method as a set of steps and tools to be applied to certain problems, but also on how this protocol is introduced to prospective users, how these users are trained, what documentation they are provided with, and what incentives they are given to engage with the new method. The intervention that needs evaluating is this full package, and not only the method. This supposes complementing the description of design methods by Gericke et al. (2017), which stops at the level of the "intended use" of the methods, with a description of how this "intended use" is made to happen in a given situation (or set of situations).
To achieve a good understanding of how methods could generate impact, "programme theories" could borrow from theories beyond the field of engineering design. Design phenomena are studied in a wide range of disciplines (Cash, 2018), and an even broader range of theories can be used to explain why certain methods are adopted and generate benefits, while others do not. In health sciences, intervention designers and evaluators have sometimes borrowed from behavioural, social, organisational and cognitive sciences to support the design and evaluation of interventions. This allows cumulative learning across disciplines and could support better translation of the design knowledge embedded in design methods into everyday design practice, through improved implementation mechanisms.

4. Implications for engineering design methods

In the following, we summarize and discuss potential implications of evaluation approaches from software engineering and health sciences. As mentioned in the introduction, the underlying assumption is that engineering design, software engineering, and health sciences all use methods for generating a desired outcome in an organizational context. At the same time, we acknowledge the fundamental differences between these domains, which may limit the implications for engineering design. Software engineering focuses on the design of software. For health sciences, the objective is ultimately to improve treatment outcome. Hence, the object of design is different from engineering

design and the implications are limited to the evaluation of the design process rather than the object of design.
Table 1 provides an overview of how far evaluation approaches from the two domains studied have already been applied for evaluating design methods, how far they are pertinent for evaluating design methods, and challenges. As mentioned in Section 2, most software engineering evaluation approaches (case studies, surveys, experiments, ethnographies, action research) are used in engineering design as well. They might only differ in the number of studies conducted for evaluating a specific design method. What could be learned from software engineering is how to raise awareness of the lack of evidence for the effect of design methods, which is best remedied via a research program focusing on empirical methods.

Table 1. Overview of evaluation practices in software engineering, health and medical sciences and their potential implications for evaluating design methods

Evaluation approach	Already applied to the evaluation of design methods?	Pertinent to evaluate design methods?	Challenges
Evaluation of software engineering methods	Yes	Yes	Challenges are similar in both fields
Evaluation of complex healthcare interventions			
- Experimental and semi-experimental evaluation	Rare, often no control group or no consideration of the counterfactual (what is most likely to have happened if the design method had not been used?).	Depending on the method, can be practically challenging.	Difficulty of defining outcomes and indicators for quantitative comparison Difficulty of implementing similar method across multiple settings for comparison Lack of awareness and training in evaluation methods
- Theory-driven evaluation	Partly proposed in DRM (Blessing and Chakrabarti, 2009), but rarely seen in publications.	Yes. Congruent with recent calls for theory-driven design research.	Identifying relevant theoretical corpora. Awareness of the approach.
- Full description of intervention systems	Often incomplete account of how implementation strategy of design methods might affect outcomes	Yes	Awareness of the requirement.
Both			
- Meta-analyses	Rare, starting to be published (Cash, 2018).	Yes	Lack of agreement on unified evaluation methods across studies.

Regarding the evaluation of complex healthcare interventions, they have been rarely or only partly applied to design methods. Experimental and semi-experimental evaluations have rarely been applied to design methods in practical settings for obvious reasons such as the difficulty of using a control group due to resource constraints, e.g. two teams working on a sufficiently complex design project with similarly qualified team members. However, beyond the lack of resources of conducting such a study, the field of engineering design lacks clearly defined outcomes and indicators which are accepted in the community and used across studies. The lack of such standardized outcomes and indicators prevents the comparison between studies. Finally, there is a lack of awareness and training of evaluation methods. These points are almost identical to those mentioned in Sjoberg et al. (2007)

for the field of software engineering. Both domains seem to face similar issues, however, the field of software engineering has debated these issues already about 10 to 15 years ago.
Theory-driven evaluation was proposed in Blessing and Chakrabarti (2009). However, it seems to be rarely in use (Lamé, 2019). Nevertheless, Cash (2018) argues for the importance of theory-driven design research. A challenge for a theory-driven research agenda in engineering design is the identification of relevant theoretical corpora.
The full description of intervention systems has only partly been adopted in engineering design. The main issue is that the link between the intervention strategy and the outcome is incompletely described, which makes it difficult to link cause and effect. Increasing the rigour of such studies could lead to more relevant results.
Regarding approaches used in both fields, meta-analyses are clearly of interest, due to their ability to gain insights into the effectiveness of methods across studies. Meta-analyses are rare in engineering design (Lamé, 2019), due to the lack of comparable studies for specific design methods. The application of meta-analyses to engineering design requires the satisfaction of preconditions such as: 1) Sufficient number of studies of the same type; 2) Same hypothesis across studies; 3) Common quantitative measures for explanatory variables, controls, and effect size across studies; 4) Avoidance of publication bias (bias towards publications which report statistically significant differences). These conditions are currently not satisfied by the vast majority of evaluation approaches in engineering design.
For making at least some minimal progress towards applying the above-mentioned approaches to engineering design, we propose the following steps:

- Development of empirical data bases where qualitative and quantitative results from empirical design research are stored according to standardized criteria;
- Development of agreed upon theoretical constructs such as causal graphs which can be explored empirically in a research program;
- Development of agreed upon measures for empirical studies, which allows for comparability between studies;
- Collaboration for maximizing the number of studies on specific design methods.

It is clear that implementing these steps is challenging and will take years. It also requires a level of collaboration and alignment which seems to be difficult to achieve. Nevertheless, we agree with Cash (2018) that "without action to increase scientific, theoretical, and methodological rigour there is a real possibility of the field being superseded and becoming obsolete through lack of impact."

5. Conclusions

In this paper we provide perspectives on how the evaluation of engineering design methods could benefit from transferring validation approaches from the health sciences and software engineering. For this purpose, we survey approaches from these domains and show which approaches could be transferred to engineering design methods. Several approaches seem to be pertinent for design methods such as (semi-) experimental evaluation, theory-driven evaluation, full description of intervention systems, and meta-analyses. However, real progress on evaluating design methods can only be expected if preconditions such as standardized theoretical constructs, measures, data bases of empirical data, and a sufficient number of studies on specific design methods are developed. For future work, we propose the identification of specific design methods for which the collection of substantial empirical evidence would be at the same time feasible and expected to have significant practical impact.

References

Badke-Schaub, P. and Frankenberger, E. (1999), "Analysis of design projects", *Design Studies*, Elsevier, Vol. 20 No. 5, pp. 465-480.
Barth, A., Caillaud, E. and Rose, B. (2011), "How to validate research in engineering design?", *ICED 2011*, Vol. Vol. 2: De, The Design Society, Lyngby/Copenhagen.
Blessing, L. and Chakrabarti, A. (2009), *DRM, a Design Research Methodology*, Springer, London.

Breuer, E. et al. (2016), "Using theory of change to design and evaluate public health interventions: a systematic review", *Implementation Science : IS*, Vol. 11, p. 63.
Bryant, C.R. et al. (2006), "A validation study of an automated concept generator design tool", *ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pp. 283-294.
Bunning, C. (1995), *Professional Development Using Action Research. Look Forward Ask Questions Learn Virtual Conference*, MCB University Press.
Cash, P.J. (2018), "Developing theory-driven design research", *Design Studies*, Vol. 56, pp. 84-119.
Craig, P. et al. (2008), "Developing and evaluating complex interventions: the new Medical Research Council guidance", *BMJ (Clinical Research Ed.)*, Vol. 337, pp. a1655-a1655.
Cristiano, J.J., Liker, J.K. and White, C.I. (2001), "Key factors in the successful application of quality function deployment (QFD)", *IEEE Transactions on Engineering Management*, Vol. 48 No. 1, pp. 81-95.
Dingsøyr, T., Dybå, T. and Abrahamsson, P. (2008), "A preliminary roadmap for empirical research on agile software development", *Agile 2008 Conference*, IEEE, pp. 83-94.
Dorst, K. (2008), "Design research: a revolution-waiting-to-happen", *Design Studies, Elsevier*, Vol. 29 No. 1, pp. 4–11.
Dybå, T. and Dingsøyr, T. (2008), "Strength of evidence in systematic reviews in software engineering", *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ACM, pp. 178-187.
Dybå, T., Sjøberg, D.I. and Cruzes, D.S. (2012), "What works for whom, where, when, and why?: on the role of context in empirical software engineering", *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ACM, pp. 19-28.
Easterbrook, S. et al. (2008), "Selecting empirical methods for software engineering research", *Guide to Advanced Empirical Software Engineering*, Springer, pp. 285-311.
Fenton, N. (2001), "Viewpoint Article: Conducting and presenting empirical software engineering", *Empirical Software Engineering*, Vol. 6 No. 3, pp. 195-200.
Ferreira, I.M. and Gil, P.J. (2012), "Application and performance analysis of neural networks for decision support in conceptual design", *Expert Systems with Applications*, Vol. 39 No. 9, pp. 7701-7708.
Frey, D.D. and Dym, C.L. (2006), "Validation of design methods: lessons from medicine", *Research in Engineering Design*, Vol. 17 No. 1, pp. 45-57.
Gericke, K., Eckert, C.M. and Stacey, M. (2017), "What do we need to say about a design method?", *ICED 2017*, Vol. 7: De, The Design Society, Vancouver, BC, Canada.
Green, J. and Britten, N. (1998), "Qualitative research and evidence based medicine", *BMJ*, Vol. 316 No. 7139, pp. 1230-1232.
Griffin, A. (1991), *Evaluating Development Processes: QFD as an Example*.
Hannay, J.E. et al. (2009), "The effectiveness of pair programming: A meta-analysis", *Information and Software Technology*, Vol. 51 No. 7, pp. 1110-1122.
Hawe, P., Shiell, A. and Riley, T. (2009), "Theorising interventions as events in systems", *American Journal of Community Psychology*, Vol. 43 No. 3-4, pp. 267-276.
Hedges, L.V. and Olkin, I. (2014), *Statistical Methods for Meta-Analysis*, Academic press.
Hein, A. (2016), *Heritage Technologies in Space Programs - Assessment Methodology and Statistical Analysis*, PhD thesis, Technical University of Munich.
Kitchenham, B.A. et al. (2002), "Preliminary guidelines for empirical research in software engineering", *IEEE Transactions on Software Engineering*, Vol. 28 No. 8, pp. 721-734.
Ko, A.J., Latoza, T.D. and Burnett, M.M. (2015), "A practical guide to controlled experiments of software engineering tools with human participants", *Empirical Software Engineering*, Vol. 20 No. 1, pp. 110-141.
Lamé, G. (2019), "Systematic literature reviews: an introduction", *ICED19 - 22nd International Conference on Engineering Design*, Design Society.
Lamé, G. and Dixon-Woods, M. (n.d.). "Using clinical simulation to study how to improve quality and safety in healthcare", *BMJ Simulation and Technology Enhanced Learning*, available at: https://doi.org/10.1136/bmjstel-2018-000370.
Merlin, T., Weston, A. and Tooher, R. (2009), "Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'", *BMC Medical Research Methodology*, Vol. 9 No. 1, p. 34.
Minary, L. et al. (2019), "Which design to evaluate complex interventions? Toward a methodological framework through a systematic review", *BMC Medical Research Methodology*, Vol. 19 No. 1, p. 92.
Moore, G.F. et al. (2015), "Process evaluation of complex interventions: Medical Research Council guidance", *BMJ*, Vol. 350, available at: https://doi.org/10.1136/bmj.h1258.
Murad, M.H. et al. (2016), "New evidence pyramid", *Evidence-Based Medicine*, Vol. 21 No. 4, pp. 125-127.

Perry, D.E., Porter, A.A. and Votta, L.G. (2000), "Empirical studies of software engineering: a roadmap", *Proceedings of the Conference on The Future of Software Engineering*, ACM, pp. 345-355.
Pickard, L.M., Kitchenham, B.A. and Jones, P.W. (1998), "Combining empirical results in software engineering", *Information and Software Technology*, Vol. 40 No. 14, pp. 811-821.
Radcliffe, D. and Harrison, P. (1994), *Transforming Design Practice in a Small Manufacturing Enterprise, American Society of Mechanical Engineers, Design Engineering Division, Design Engineering Division* (Publication) DE.
Rafique, Y. and Mišić, V.B. (2012), "The effects of test-driven development on external quality and productivity: A meta-analysis", *IEEE Transactions on Software Engineering*, Vol. 39 No. 6, pp. 835-856.
Runeson, P. et al. (2012), *Case Study Research in Software Engineering: Guidelines and Examples*, John Wiley & Sons.
Seepersad, C.C. et al. (2006), "The Validation Square: How Does One Verify and Validate a Design Method?", in Lewis, K.E., Chen, W. and Schmidt, L.C. (Eds.), *Decision Making in Engineering Design*, ASME, New York, NY, available at: http://doi.org/10.1115/1.802469.ch25.
Shi, Y.L.Z. et al. (2019), "Cognitive Style and Field Knowledge in Complex Design Problem-Solving: A Comparative Case Study of Decision Support Systems", *Design Computing and Cognition '18*, Springer International Publishing, pp. 341-360.
Shull, F., Singer, J. and Sjøberg, D.I. (2007), *Guide to Advanced Empirical Software Engineering*, Springer Science & Business Media.
De Silva, M.J. et al. (2014), "Theory of Change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions", *Trials*, Vol. 15 No. 1, p. 267.
Sio, U.N., Kotovsky, K. and Cagan, J. (2015), "Fixation or inspiration? A meta-analytic review of the role of examples on design processes", *Design Studies*, Vol. 39, pp. 70-99.
Sjoberg, D.I., Dyba, T. and Jorgensen, M. (2007), "The future of empirical methods in software engineering research", *2007 Future of Software Engineering*, IEEE Computer Society, pp. 358-378.