FEDS

A Framework for Evaluation in Design Science Research

Venable, John; Pries-Heje, Jan; Baskerville, Richard

Published in:
European Journal of Information Systems

DOI:
10.1057/ejis.2014.36

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, *25*(1), 77-89. https://doi.org/10.1057/ejis.2014.36

European Journal of Information Systems (2016) 25, 77–89

www.palgrave-journals.com/ejis/

RESEARCH ESSAY

Open

FEDS: a Framework for Evaluation in Design Science Research

John Venable[1], Jan Pries-Heje[2] and Richard Baskerville[1,3]

[1] *School of Information Systems, Curtin University, Perth, Western Australia, Australia;* [2] *Roskilde University, Roskilde, Denmark;* [3] *Georgia State University, Atlanta, Georgia, USA*

Correspondence: John Venable, School of Information Systems, Curtin University, GPO Box U1987, Perth, WA 6845, Australia.
Tel: +61 8 9266 7054;
Fax: +61 8 9266 3076;
E-mail: j.venable@curtin.edu.au

Abstract
Evaluation of design artefacts and design theories is a key activity in Design Science Research (DSR), as it provides feedback for further development and (if done correctly) assures the rigour of the research. However, the extant DSR literature provides insufficient guidance on evaluation to enable Design Science Researchers to effectively design and incorporate evaluation activities into a DSR project that can achieve DSR goals and objectives. To address this research gap, this research paper develops, explicates, and provides evidence for the utility of a Framework for Evaluation in Design Science (FEDS) together with a process to guide design science researchers in developing a strategy for evaluating the artefacts they develop within a DSR project. A FEDS strategy considers why, when, how, and what to evaluate. FEDS includes a two-dimensional characterisation of DSR evaluation episodes (particular evaluations), with one dimension being the functional purpose of the evaluation (formative or summative) and the other dimension being the paradigm of the evaluation (artificial or naturalistic). The FEDS evaluation design process is comprised of four steps: (1) explicate the goals of the evaluation, (2) choose the evaluation strategy or strategies, (3) determine the properties to evaluate, and (4) design the individual evaluation episode(s). The paper illustrates the framework with two examples and provides evidence of its utility via a naturalistic, summative evaluation through its use on an actual DSR project.
European Journal of Information Systems (2016) **25**(1), 77–89.
doi:10.1057/ejis.2014.36; published online 11 November 2014

Keywords: Design Science Research; research methodology; information systems evaluation; utility evaluation; artefact evaluation; research design

Introduction

Evaluation of design artefacts and design theories is a central and critical part of Design Science Research (DSR) (March & Smith, 1995; Hevner *et al*, 2004; Vaishnavi & Kuechler, 2004). In DSR, evaluation is primarily concerned with evaluation of design science outputs, including Information Systems (IS) Design Theories (Gregor & Jones, 2007) and design artefacts (March & Smith, 1995). Together with 'build', evaluation is one of two key activities that constitute DSR (March & Smith, 1995). As other research paradigms (positivist, interpretivist, critical) do not design, develop, or 'build' new artefacts (else they would be DSR), design artefact and design theory evaluation are much more relevant, important, and specific to DSR than to other research paradigms. Evaluation is 'crucial' to DSR and requires researchers to rigorously demonstrate the utility, quality, and efficacy of a design artefact using well-executed evaluation methods (Hevner *et al*, 2004, pp. 82, 85). Designed artefacts must be analysed as to their use and performance as possible explanations for changes (and hopefully improvements) in the behaviour of systems, people, and organisations (Vaishnavi & Kuechler, 2004).

Received: 28 October 2012
Revised: 14 October 2013
2nd Revision: 30 June 2014
Accepted: 26 August 2014

As part of the design science process, evaluation may be tightly coupled with design itself. This tight linkage arises from the impact of evaluations on designer thinking, with the potentially rapid cycles of build and evaluate that sometimes constitute design itself.

Without sound evaluation, DSR must conclude with only theorising about the utility of design artefacts, that is, with an assertion that a new technology 'works' without any evidence that it does. Because its context includes research goals, evaluation in DSR has a broader purpose than in the 'ordinary' practice of design. In an ordinary design project without scientific aims, evaluation is focused on evaluating the artefact in the context of the utility it contributes to its environment (Hevner *et al*, 2004, call this the relevance cycle). In a design science project, evaluation must also regard the design and the artefact in the context of the knowledge it contributes to the knowledge base (Hevner *et al*, 2004 call this the rigour cycle). Since such a build-and-evaluate cycle seeks to deliver both environmental utility and additional (new) knowledge, the evaluation approach not only needs to address the quality of the artefact utility, but also the quality of its knowledge outcomes. This dual purpose of evaluation means that, if DSR is to live up to its label as 'science', the evaluation should be relevant, rigorous, and scientific.

But how should such evaluations be designed and conducted as part of a DSR project? What strategies and methods should be used for evaluation in a particular DSR project? How can the evaluation be designed to be both effective (rigorous) and efficient (prudently using resources, including time)? The extant DSR literature identifies a variety of different evaluation methods (e.g., in Nunamaker *et al*, 1990/1991; March & Smith, 1995; Hevner *et al*, 2004; Vaishnavi & Kuechler, 2004; Venable, 2006; Peffers *et al*, 2008; Gill & Hevner, 2013), but provides precious little guidance for deciding what to evaluate and which evaluation methods to use or why, when, and how to use them to best conduct the evaluation component(s) of a DSR project or programme. Moreover, much of this existing literature assumes that only one kind of evaluation will be necessary to demonstrate both the artefact's utility, fitness, or usefulness (*cf*. Gill & Hevner, 2013), as well as any design principles or theory employed in the artefact's construction. Furthermore, the cyclical nature of many design science processes may demand different evaluations at different stages of progress. These gaps give rise to the following research question:

'What would be a good way to guide the design of an appropriate strategy for conducting the various evaluation activities needed throughout a DSR project?'

The research reported in this paper extends the authors' earlier work (Pries-Heje *et al*, 2008; Venable *et al*, 2012) to answer the above research question by (1) developing and (2) evaluating the utility of (a) a new, enhanced Framework for Evaluation in Design Science (FEDS), which revises the earlier framework dimensions and adds new concepts of evaluation strategies and trajectories, together with (b) a new evaluation design process for applying that framework. The FEDS framework and evaluation design process together can be used to support and guide DSR researchers (especially novice researchers) in the design of the evaluation component(s) of their DSR projects and programmes. The paper motivates, proposes, illustrates, and presents evaluations of the FEDS framework and evaluation design process.

The next section reviews selected literature on evaluation to further develop the research gap and research question introduced above and to provide a basis to inform the design of the FEDS Framework and Evaluation Design Process. Next, the subsequent two sections articulate the form and rationale for the FEDS Framework and Evaluation Design Process proposed in this paper. Following that, the subsequent two sections illustrate the framework by applying it to two examples from the DSR literature and describe a naturalistic summative evaluation of FEDS (through its use to guide the evaluation strategy choices made in a particular DSR project) to provide evidence of the utility of FEDS in practice. Finally, the last two sections discuss the findings, summarise the research, identify limitations, and give some suggestions for further research.

Evaluation in the literature

Remenyi (1999) identifies the two most important categories of evaluation as (1) formative *vs* summative evaluation and (2) *ex ante vs ex post* evaluation (Smithson & Hirschheim, 1998; Stefanou, 2001; Irani & Love, 2002; Klecun & Cornford, 2005). Other evaluation categories regard (3) the distinctions between approaches and techniques, for example, quantitative *vs* qualitative approaches or subjective *vs* objective techniques (Remenyi, 1999). Such categories distinguish the reasoning and strategies for evaluation including (1) *why* to evaluate, (2) *when* to evaluate, and (3) *how* to evaluate. Another important aspect is (4) *what* to evaluate: the properties of the evaluand to be examined during an evaluation (Stufflebeam, 2003).

Why to evaluate: formative *vs* summative evaluation

The main distinction in why to evaluate is formative *vs* summative evaluation. This distinction does not arise in the innate qualities of the evaluation process, but rather inhabits the functional purpose of the evaluation. For example, an evaluation process that may have been formulated for summative purposes may also be put to use for formative purposes (William & Black, 1996).

Formative evaluations are used to produce empirically based interpretations that provide a basis for successful action in improving the characteristics or performance of the evaluand. Formative evaluations focus on consequences and support the kinds of decisions that intend to improve the evaluand (William & Black, 1996).

Summative evaluations are used to produce empirically based interpretations that provide a basis for creating shared meanings about the evaluand in the face of different contexts. Summative evaluations focus on meanings and support the kinds of decisions that intend to influence

the selection of the evaluand for an application (William & Black, 1996).

When to evaluate: *ex ante vs ex post* evaluation

The main distinction in when to evaluate is *ex ante vs ex post* evaluation. This distinction arises from the timing of the evaluation episodes. Formative evaluation episodes are often regarded as iterative or cyclical (William & Black, 1996) in order to measure improvement as development progresses. Summative evaluation episodes are more often used to measure the results of a completed development or to appraise a situation before development begins.

Ex-ante evaluation is 'the predictive evaluation which is performed in order to estimate and evaluate the impact of future situations' (Stefanou, 2001, p. 206). For IS development, *ex ante* evaluation serves the purpose of deciding whether or not to acquire or develop a technology, or the purpose of deciding which of several competing technologies should be acquired or adopted. It happens before design and construction begins.

Ex post evaluation is an assessment of 'the value of the implemented system on the basis of both financial and non-financial measures' (Stefanou, 2001, p. 206). Approaches to *ex post* evaluation in IS can be derived from Symons' (1991) critical adaptation of the 'context, content and process' model developed for organisational change evaluation. Examples of variations include interpretive (Stockdale & Standing, 2006) and critical (Klecun & Cornford, 2005) evaluation approaches.

In terms of timing, *ex ante* and *ex post* evaluations occupy the two extremes of an evaluation continuum, as shown in Figure 1. An *ex ante* evaluation regards candidate systems or technologies before they are chosen, acquired, implemented, designed, or constructed. An *ex post* evaluation regards a chosen and developed system or technology after it has been acquired, designed, constructed, or implemented (Klecun & Cornford, 2005). Figure 1 illustrates how evaluations can also occur intermediately between *ex ante* and *ex post*. It may seem intuitive that *ex post* evaluations are always summative and *ex ante* and intermediate evaluations are always formative. However, *ex ante* and *ex post* refer only to timing. A summative evaluation may be required on an *ex ante* or intermediate basis (e.g., for continuation approval) and *ex post* evaluations may also have formative purposes.

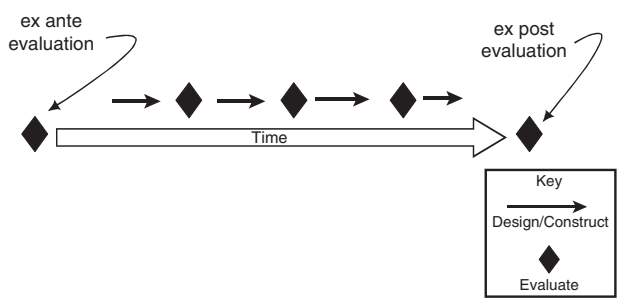

Figure 1 *Ex ante–ex post* evaluation time continuum.

The distinctions of formative *vs* summative and *ex ante vs ex post* within the IS literature is focused on a *particular* system or technology to address a *particular, situated* problem. For the purpose of evaluation in DSR, we translate these concepts to address evaluating a new *kind* of artefact for addressing a *kind* of problem.

Why to evaluate: purpose and goals of evaluation in DSR

A review of the DSR literature elaborates not just two, but (at least) six different (but related) purposes for the evaluation activity in DSR. First, one key purpose of evaluation in DSR is to determine how well a designed artefact or ensemble of artefacts achieves its expected environmental utility (an artefact's main purpose).

A second key purpose of evaluation is the substantiation of design theory in terms of the quality of the knowledge outcomes (Baskerville *et al*, 2007; Kuechler & Vaishnavi, 2012), that is, to provide evidence that the theory leads to some developed artefact that will be useful for solving some problem or making some improvement.

Third, evaluation may also be concerned with comparative evaluation of the new artefact (or design theory) in comparison with other artefacts (or design theories) (Venable, 2006) to determine whether the new artefact/design theory makes an improvement on the state of the art.

A fourth purpose considers that utility is a complex, composite concept, which is composed of a number of different criteria beyond simple achievement of an artefact's main purpose. Together with style, the 'utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods. … artifacts can be evaluated in terms of functionality, completeness, consistency, accuracy, performance, reliability, usability, fit with the organization, and other relevant quality attributes' (Hevner *et al*, 2004, p. 85).

Fifth, an artefact may be evaluated 'for other (undesirable) impacts' (Venable, 2006), otherwise known as side effects.

Sixth and finally, evaluation can further elaborate the knowledge outcomes by discerning *why* an artefact works or not (Vaishnavi & Kuechler, 2004).

Problems with evaluation

Evaluation in DSR is potentially fraught with problems. Potential errors include Type I and Type II errors, otherwise known as a false positive and a false negative (Baskerville *et al*, 2007). In DSR, a false positive is a finding that a new artefact works (or its corresponding design theory is correct) when in fact the artefact does not work (or its corresponding design theory is incorrect). A false negative is a finding that a new artefact does not work (or its corresponding design theory is incorrect) when in fact the artefact does work (or its corresponding design theory is correct). Baskerville *et al* (2007) analyse the evaluation process to identify a number of potential sources of errors, which they suggest be considered when designing and carrying out an evaluation in DSR. While they suggest the use of qualitative approaches for data gathering and

analysis for evaluation, they do not provide much further guidance. It also is not clear whether what they propose is appropriate for all kinds of artefacts.

Summary of DSR evaluation literature

The above literature identifies a number of different purposes for evaluation in DSR, as well as a variety of different evaluation paradigms, methods, and activities. However, there is little or no guidance provided in how or why a DSR researcher can or should choose among the different paradigms or methods to achieve a DSR project's evaluation goals. This research gap in DSR has motivated this paper and the research question stated in the introduction, which is: 'What would be a good way to guide the design of an appropriate strategy for conducting the various evaluation activities needed throughout a DSR project?'

We draw and build upon the ideas from the above literature in the next section: The FEDS Framework for Evaluation in Design Science, which has the goal of helping to specifically guide DSR researchers in the design of an appropriate strategy and evaluation activities according to the needs of their DSR project or programme.

The FEDS Framework for Evaluation in Design Science Research

As noted in the introduction, FEDS is designed to address the research question 'What would be a good way to guide the design of an appropriate strategy for conducting the various evaluation activities needed throughout a DSR project?' FEDS was designed to help DSR researchers, especially novices, decide on an appropriate strategy or strategies for evaluating the outcomes of the build activity in DSR. We also developed a process for using the framework in designing the particular evaluation research strategy. This section describes the FEDS framework itself, while the next section describes a process for using the framework.

We developed the FEDS framework analytically by looking at the different classifications of extant evaluation methods and relating them to the goals of evaluation in DSR. The goals are the varying objectives of evaluation while evaluation methods are the means. The framework provides a way to support evaluation research design decisions by creating a bridge between the evaluation goals and evaluation strategies. By providing a classification of evaluation strategies and relating strategies to goals, FEDS provides this bridge. Two important aspects or dimensions determined by the analysis above are (1) the functional purpose of the evaluation (formative or summative) and (2) the paradigm of the evaluation (artificial or naturalistic). These two dimensions form the basis of the FEDS framework.

Dimension 1: functional purpose of the evaluation

Why to evaluate: Formative and summative evaluations are distinguished by their functional purpose rather than any difference in the nature of the content of their evaluations. The functional purpose of formative evaluations is to help improve the outcomes of the process under evaluation. The functional purpose of summative evaluations is to judge the extent that the outcomes match expectations, for example, certification, progress, or even the effectiveness of the process itself (William & Black, 1996). Because the distinction is purposive, the mechanics of any particular evaluation activity may yield evidence that is useful both formatively and summatively. The formative and summative functional purposes of evaluations can be characterised as the ends of a continuum along which any evaluation might be located, as shown on the *x*-axis of the FEDS Framework in Figure 2. Towards the formative end, evaluations must provide a basis for successful action. Towards the summative end, evaluations must create a consistent interpretation across shared meanings (like standards or requirements). Stated simply, 'when formative functions are paramount, meanings are validated by their consequences, and when summative functions are paramount, consequences are validated by meanings' (William & Black, 1996, p. 545).

Dimension 2: paradigm of the evaluation study

How to evaluate: As introduced in the second section, a DSR evaluation method has a paradigm in a sense similar to scientific paradigms like positivism or interpretivism. While there are different ways to characterise such paradigms, the prescriptive and functional nature of design science demands a distinction that is more practical and less philosophical. For the second dimension of our framework, we adopt the distinction between artificial evaluation and naturalistic evaluation made by Venable (2006) and place it along the *y*-axis of FEDS, as shown in Figure 2.

Artificial evaluation may be empirical or non-empirical (e.g., logical/rhetorical). It is nearly always positivist and reductionist, being used to test design hypotheses (Walls *et al*, 1992). However, interpretive techniques may also be used to attempt to better understand why an artefact works or why it work. Even critical techniques may be used, but these generally supplement the main goal of

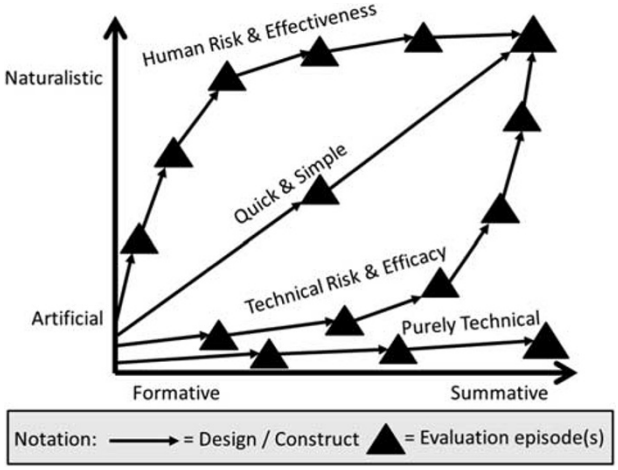

Figure 2 FEDS (Framework for Evaluation in Design Science) with evaluation strategies.

proving or disproving the design theory and/or the utility of the DSR artefacts. Artificial evaluation includes laboratory experiments, simulations, criteria-based analysis, theoretical arguments, and mathematical proofs. The dominant scientific/rational paradigm brings to artificial evaluation the benefits of stronger scientific reliability in the form of better repeatability and falsifiability (Gummesson, 1988).

Naturalistic evaluation explores the performance of a solution technology in its real environment, typically within an organisation. By performing evaluation in a real environment (i.e., real people, real systems, and real settings, Sun & Kantor, 2006), naturalistic evaluation embraces all of the complexities of human practice in real organisations. Naturalistic evaluation is always empirical and tends towards interpretivism, but may be positivist and/or critical. Naturalistic evaluation methods typically include case studies, field studies, field experiments, surveys, ethnography, phenomenology, hermeneutic methods, and action research. The dominant interpretive paradigm brings to naturalistic DSR evaluation the benefits of stronger internal validity (Gummesson, 1988).

On the one hand, artificial evaluation is often the simplest, most straightforward, and least costly form of evaluation. It often affords very precise language in its findings. Since it usually controls for the obvious confounding variables, it is less susceptible to misinterpretation and bias. Naturalistic evaluation can be difficult (and costly), partly because it must disentangle the effects of many confounding variables in a real world setting. To the extent that naturalistic evaluation is affected by confounding variables or misinterpretation, evaluation results may not be precise or even truthful about an artefact's utility or efficacy in real use.

On the other hand, artificial evaluation involves reductionist abstraction from the natural setting (in order to assure rigour in its assessment of efficacy of the technology artefact) and is necessarily unrealistic in the sense that it fails to adhere to one or more of the three realities (i.e., unreal users, unreal systems, or unreal problems) of Sun & Kantor (2006). To the extent that an artificial evaluation setting is unreal, evaluation results may not correspond to real use. In contrast, naturalistic evaluation offers more critical face validity and also assures more rigorous assessment of the effectiveness of the artefact.

The two dimensions (that can be seen as the *x*- and *y*-axis in Figure 2) are fully orthogonal to each other. Both naturalistic and artificial evaluation methods can be used for formative and/or summative evaluations, with the advantages and disadvantages described above.

Generally evaluation progresses from a state of no evaluation having been conducted at the origin towards a more comprehensive and rigorous (in the sense of evaluating more fully and realistically) in the upper right corner. The chronological progression through formative evaluations to more summative evaluation represents the purpose, peculiar to DSR, to rigorously consider the quality of the knowledge outcomes. The increasing use of more summative evaluations enables comparison of research outcomes with research expectations (testing the design theory). The chronological progress through artificial evaluations to more naturalistic evaluation represents a similar, but subtly different purpose. The increasing use of more naturalistic evaluations improves the quality of the knowledge outcomes *concerning the artefact's effectiveness in real use*, as the artefact increases in quality and the risks become low enough for real use by real users.

While evaluation typically progresses from the lower left to the upper right of Figure 2, there are many paths or trajectories that may be followed in conducting a number of evaluation episodes, that is, specific evaluation activities of specific evaluands using a specific evaluation method. A planned trajectory of evaluations that is appropriate for the circumstances of a particular DSR project is an *evaluation strategy*. But what different evaluation strategies are there and how should one choose one? The next section answers this question by showing prototypical strategies and discussing why one would want to choose one (or more) of them.

Evaluation strategies

When to evaluate, for what purpose, and how: The pathway or trajectory sought and followed in a DSR project or programme may differ according to the needs and resources available to the DSR project/programme. This gives rise to different strategies. Each strategy operates as a progression that proceeds from the origin of the evaluation framework towards some final summative evaluation that concludes the DSR project or programme. Our analysis identified four different possible strategies (others may be possible), as shown in Figure 2. The strategies we identified include the Quick & Simple strategy, the Human Risk & Effectiveness evaluation strategy, the Technical Risk & Efficacy evaluation strategy, and the Purely Technical Artefact strategy. The triangles in Figure 2 show evaluation episodes or where the evaluations occur in the strategy. The number of triangles and their placement along any particular strategy's trajectory in Figure 2 are indicative only; they may (and should) vary according to the needs of a particular DSR project/programme. Furthermore, any planned strategy may need to be revised during the course of a particular DSR project or programme.

The Quick & Simple strategy conducts relatively little formative evaluation and progresses quickly to summative and more naturalistic evaluations. The evaluation trajectory of this strategy includes relatively few evaluation episodes (perhaps even only one summative evaluation at the end). Such a strategy is low cost and encourages quick project conclusion, but may not be reasonable in the face of various design risks.

The Human Risk & Effectiveness evaluation strategy emphasises formative evaluations early in the process, possibly with artificial, formative evaluations, but progressing quickly to more naturalistic formative evaluations. Near the end of this strategy more summative evaluations are engaged, which focus on rigorous evaluation of the effectiveness of the artefact, that is, that the utility/benefits

of the artefact will continue to accrue even when the artefact is placed in operation in real organisational situations and over the long run, despite the complications of human and social difficulties of adoption and use.

The Technical Risk & Efficacy evaluation strategy emphasises artificial formative evaluations iteratively early in the process, but progressively moving towards summative artificial evaluations. Artificial summative evaluations are used to rigorously determine efficacy of the artefact, that is, that the utility/benefits derived from the use of the artefact are due to the artefact, not due to other factors. Near the end of this strategy more naturalistic evaluations are engaged.

A fourth strategy, the Purely Technical strategy is used when an artefact is purely technical, without human users, or planned deployment with users is so far removed from what is developed to make naturalistic evaluation irrelevant. This strategy is similar to the Quick & Simple strategy, but favours artificial over naturalistic evaluations throughout the process, as naturalistic strategies are irrelevant to purely technical artefacts or when planned deployment with users is far in the future.

Table 1 summarises the relevant circumstances when we might select each of the four strategies.

The evaluation strategies described above are prototypical ways to address particular goals. However, it is possible that multiple, different goals might call for more than one strategy or a combined, hybrid evaluation strategy. For example, a new kind of hospital technology might be developed that would notify an emergency room physician of a patient's need for urgent care along with notifying the patient's usual physician – with a goal of generating interaction between the two physicians. Two different goals are involved: one is the operating utility of the new technical artefact and the other is the utility in the new social behaviour. A hybrid evaluation strategy might begin with a purely technical strategy in order to develop knowledge about the new kind of artefact itself. Once the hospital artefact is proved operational, the evaluation strategy might shift to one of human risk and effectiveness in order to develop knowledge about new kinds of social behaviour among the physicians.

This example also illustrates the need for a DSR-project-specific evaluation strategy. Unlike a design setting in which the only goals are the utility of the artefact and its effects on social behaviour, such a DSR setting adds further goals for contributing rigorously developed knowledge about this new kind of artefact and its effects on its environment.

Table 1 Circumstances for selecting a relevant DSR evaluation strategy

DSR evaluation strategies	*Circumstance selection criteria*
Quick & Simple	If small and simple construction of design, with low social and technical risk and uncertainty
Human Risk & Effectiveness	If the major design risk is social or user oriented *and/or* If it is relatively cheap to evaluate with real users in their real context *and/or* If a critical goal of the evaluation is to rigorously establish that the utility/benefit will continue in real situations and over the long run
Technical Risk & Efficacy	If the major design risk is technically oriented *and/or* If it is prohibitively expensive to evaluate with real users and real systems in the real setting and/or If a critical goal of the evaluation is to rigorously establish that the utility/benefit is due to the artefact, not something else
Purely Technical Artefact	If artefact is purely technical (no social aspects) or artefact use will be well in future and not today

An evaluation strategy choice process for DSR

On the basis of the framework above we can derive a four-step process for choosing an approach for a particular DSR project. The four steps we propose are: (1) explicate the goals of the evaluation, (2) choose the evaluation strategy or strategies, (3) determine the properties to evaluate, and (4) design the individual evaluation episode(s).

Step 1: explicate the goals

There are at least four possibly competing goals in designing the evaluation component of DSR. Some goals are more relevant at different stages of a DSR project.

Rigour: Rigour in DSR has two senses. The first sense is in establishing that it is the artefact instantiation that causes an observed outcome and only the artefact, not some confounding independent variable or circumstance (efficacy). The second is in establishing that the artefact instantiation works in a real situation (effectiveness). Artificial evaluation will likely be most appropriate for rigorously evaluating the former, while naturalistic evaluation will likely be most appropriate for rigorously evaluating the latter. Summative evaluation provides the greatest rigour in the evaluation and hence the reliability of the knowledge developed. Summative evaluations usually (but not always) occur at the end of an evaluation trajectory or strategy, that is, with (an) evaluation episode(s) towards the end of the arrows and the larger triangles in Figure 2. Possibly more than one summative evaluation episode may be required to evaluate different artefacts or their aspects or to provide stronger evidence (e.g., of their utility in different contexts).

Uncertainty and risk reduction: Formative evaluation is particularly important when design uncertainties are significant and is a key way to reduce risks due to design uncertainties. As discussed earlier, risks may be identified as human social/use risks (i.e., risks that the artefact will not fit well into the use or social situation and therefore not work or cause further problems) and technical risks

(i.e., risks that the technology cannot be made to function). Formative evaluations should be conducted as early as practicable in an evaluation trajectory or strategy, that is, with (an) evaluation episode(s) towards the beginning of the arrows in Figure 2. Identifying difficulties and areas for improvement as early as possible, so as to influence and improve the design of the artefact supports development of a higher quality (more effective, efficient, etc.) artefact and also reduces costs by resolving uncertainties and risks earlier.

Ethics: Especially in the evaluation of safety critical systems and technologies, the evaluation should address potential risks to animals, people, organisations, or the public, including future generations. More rigorous evaluations may become a goal depending on such risks. In addition, the evaluation activity itself should not put stakeholders at risk. Formative evaluation may reduce later risks, both during the evaluation to research participants and after evaluation to users of the research results. However, summative evaluation (perhaps in combination with formative evaluation) is the best way to ensure the rigour that reduces risk to the eventual users of the artefacts and knowledge resulting from DSR.

Efficiency: Efficient evaluation balances the above goals of the evaluation against the resources available for the evaluation (e.g., time and money). Formative evaluation of design artefacts can reduce costs by evaluating before incurring the costs of instantiation and theory specification in a prudent way. In general, naturalistic evaluation takes longer and will be more costly than artificial evaluation. Specific methods of evaluation are also less costly, with non-empirical (which are artificial) evaluation methods often having large savings.

Step 2: choose a strategy or strategies for the evaluation

On the basis of the goals of the evaluation, one or more strategies may be more appropriate for the evaluation. Figure 2 and Table 1 describe the strategies and when we might select each of them. Each strategy implies a decision about why, when, and how to evaluate. For example, for a socio-technical artefact with major uncertainties about social and use issues, but also with a strong need to rigorously establish long-term effectiveness in real use, the Human Risk & Effectiveness strategy would be most appropriate. In particular, at this step, we should consider the following heuristics for choosing an evaluation strategy.

(1) Evaluate and prioritise design risks, understood as potential problems that the design may face. If the major design risk is social or user-oriented, for example, related to whether the design fulfils a need or solves a problem, then pursue a Human Risk & Effectiveness strategy. If the major design risk is technically oriented, for example, whether a specific technology may work as perceived in the design, then pursue a Technical Risk & Efficacy strategy, for example, start with a laboratory experiment to clarify the boundaries of the technology.
(2) Evaluate how costly it would be to evaluate with real users and real systems in the real setting. If it is relatively cheap to have real users in their real context (setting) then pursue a Human Risk & Effectiveness strategy. By relatively cheap we mean that it is relative to the resources available in the project. A novice design science researcher may have enough time to engage with real users and contexts, but very limited development or other resources available. In this case it may be best to evaluate the design using a simple and cheap prototype first. If on the other hand, there is enough money available but relatively little time, then a Human Risk & Effectiveness strategy may buy speed for money. For example, the research project might engage and pay for the use of a usability lab. If it is too expensive to evaluate with real users and real systems in the real setting, where costly can either mean in terms of money or in terms of health or life, then pursue a Technical Risk & Efficacy strategy.
(3) Evaluate whether the artefact being developed is purely technical (not used by or affecting people) or the need or problem addressed by the design exists today or will only be deployed relatively far in the future. If the artefact is purely technical or the need to deploy the artefact exists well in the future and not today, then the Purely Technical strategy may be best for evaluation. The rationales behind this are that there is no need for human use or that real users and a real setting are not accessible (or they do not exist). Hence naturalistic evaluation is impossible.
(4) Evaluate whether the construction of the design is small and simple or large and complex. If small and simple construction, without other risks above, then construct and go directly to the Quick & Simple strategy.

Step 3: determine the properties to evaluate

The next step in strategy formulation regards what to evaluate. It entails choosing the general set of features, goals, and requirements of the artefact (design and/or instantiation) that are to be subject to evaluation.

The detailed selection of the properties is necessarily unique to the artefact, its purpose(s), and its situation during evaluation. Each artefact, within its situation, will have idiographic practical requirements that operationalise the general design theories under consideration. Different authorities have set out a wide variety of generic goals and criteria that constitute potential evaluand properties. While an encyclopaedic review is not possible in this paper, Table 2 illustrates four examples of such generic properties available for adaptation in DSR evaluation.

The framing of evaluand properties is very much dependent on the goals of the DSR project itself. For example, if the evaluand is an invoicing system governing a warehouse pick list, the Sun & Kantor (2006) cross-evaluation model would be useful. If the evaluand is an artefact embodying a security awareness training methodology,

Table 2 Examples of possible generic artefact properties

Sun & Kantor (2006)	Stufflebeam (2003)	Mathiassen *et al* (2000, based on the ISO standard 9126)	Smithson & Hirschheim (1998)
Adapting levels of granularity	Adapting context, input, process, and product	Adapting criteria as design goals	Adapting both rationality and understanding
(1) Whether the individual item was retrieved (2) Whether the task-at-hand was completed, and (3) Whether the completed task had a valuable impact on the goals-at-hand	Context: Goals Input: Strategy Process: Work plan Product: Outcomes and side effects	Useable, Secure, Efficient, Correct, Reliable, Maintainable, Testable, Flexible, Comprehensible, Reusable, Portable, and Interoperable	Rationality-efficiency: Quality assurance Rationality-effectiveness: Cost-benefit, User satisfaction, Resource utilisation Understanding: Social action, cognitive Psychology

then the CIPP model (Stufflebeam, 2003) could provide an excellent framework for the evaluand properties of interest. If the evaluand is a specification for a complex piece of software, an adaptation of the ISO 9126 standard might provide properties for evaluation (Mathiassen *et al*, 2000). ISO 9126 is a standard for measuring quality that provides a quality model with six overall dimensions: functionality, reliability, usability, efficiency, maintainability, and portability. Under each dimension in the ISO 9126 quality model, there are two to five properties that can be used for evaluation. For example, under the maintenance dimension, we find analysability, changeability, stability, and testability, which can be used to define evaluand properties. If the evaluand is a design for, or artefact embodying, an information system, then the Smithson–Hirschheim properties framework (1998) is an example of a framework for evaluating the evaluand properties of interest.

More concretely, at this step we should consider the following heuristics for choosing evaluation properties.

(1) Frame potential evaluands. In doing so we can use Table 2 for inspiration. The outcome of the framing will be a list of potential evaluands.
(2) Align candidate evaluands with the goals explicated in Step 1. Consider each potential evaluand and ask whether and to what extent it will contribute to achieving the explicated goals.
(3) Consider the strategy chosen in Step 2. If we are following a more naturalistic strategy, your evaluands should reflect that. If we are early in the formative stage, the evaluands should reflect the risks we are trying to limit and we should aim at fewer evaluands than if we are later in a summative stage.
(4) Choose evaluands based on the above heuristics (1)–(3).

Step 4: design the individual evaluation episode(s)

Having chosen a strategy or strategies and determined what properties of the artefact to evaluate, the actual evaluations need to be designed. Considering Figure 2, this step would establish what the episodes (triangles) on the figure will entail for the particular DSR project's/programme's evaluation strategy. At this step, we should consider the following heuristics for designing the individual episode.

(1) Identify and analyse the constraints in the environment. What resources are available – time, people, budget, research site, etc.? What resources are in short supply and must be used sparingly?
(2) Prioritise the above contextual factors to determine which aspects are essential, more important, less important, nice to have, and irrelevant.
(3) Decide a plan including determination of how many evaluation episodes there will be as well as when particular evaluation episodes will be conducted and in what way. Hence the outcome is: Who? Is doing what? When?

The ordering of the four steps may easily shift and iterate depending on the situation. For example, evaluand properties may have to be considered early because these could affect the functional purpose or paradigm choices. The features or content properties for evaluation of IS designs and artefacts are too diverse to enumerate here. This diversity makes universal criteria problematic, for example, because the artefacts might be either a process or a product. We must allow for different ways to frame the relevant properties of the evaluand. For example, the evaluators may choose to focus on quality criteria. There are many different perspectives to defining the notion of *quality* (Garvin, 1987). Quality is often measured in ways that reflect differences in the quantity or state of some product attribute. Another example might focus instead on measures of *success* (DeLone & McLean, 1992).

Two illustrations of evaluation strategies in the IS DSR literature

To better understand the FEDS framework, we have selected two examples from the IS literature that illustrate

two quite different choices of evaluation strategies. Each illustration has four subsections: (1) description, (2) characterisation using FEDS, (3) analysis, and (4) suggestions for improvement, in that order. Suggestions for improvement are not meant to be critical, but to highlight what might have been improved if the framework presented in this paper had been used.

Quick & Simple example

Description: Albert *et al* (2004) developed a model called GIST (Gather-Infer-Segment-Track) that can guide the design and subsequent management of web-based systems. GIST is in itself a design product that incorporates a process: 'Gather' before 'Infer', and so on. To evaluate the GIST artefact, Albert *et al* 'observe whether the redesign of the Web site in the business organization resulted in identification of business leads …' (p. 164). The website being redesigned was in a Fortune 50 company. GIST was applied and the authors '… suggested some design improvements …' (pp. 175–176). 'This resulted in a tremendous improvement …' (p. 176) and overall 'the company considers its new Web site investment and application of GIST a huge success' (p. 178).

Characterisation using FEDS: Albert *et al* (2004) apparently followed a Quick & Simple evaluation strategy. The evaluands in this case were the website development process and the website management process. The artefacts/evaluands were evaluated summatively, after the design artefacts were developed. A naturalistic evaluation assessed the improvement made by using the methods on an existing website in a Fortune 50 company. The evaluation was naturalistic in that it was conducted using a real system (the real methods) in a real organisation facing real problems. The properties under evaluation were potential improvements. The evaluation was interpretive with unspecified informants in the company reporting GIST to be a 'huge success'.

Analysis: The Quick & Simple directly summative and naturalistic strategy used by Albert *et al* is one that potentially has a high risk of failing if the evaluation goes bad, but provides the fastest evaluation of effectiveness in real use (real users, real system, with real problems to be solved). A modest assessment of the relative efficacy of the methods is also obtained because the website is improved, presumably in comparison to whatever (unstated) method was used previously.

Suggestions for improvement: While the Quick & Simple strategy, which moves quickly towards a summative naturalistic evaluation, is high risk if the evaluation goes bad, the paper does not report other formative evaluation episodes that the authors may (or may not) have used along the way to reduce risk. For example, their strategy might have started with a formative evaluation aiming for a Human Risk & Effectiveness strategy where the risk of failure is reduced in the formative evaluation, but we have no way to know. Ultimately, the authors are able to make a strong (rigorous) statement about utility (huge success) because they have put their system to a real test. Even so, a more rigorous collection and analysis of the perceived utility from the stakeholders would have provided even stronger evidence. No evidence is provided about other properties, such as ease of use, time and effort required, or ease of learning of the artefacts.

Technical risk & Efficacy example

Description: Addressing the problem of effective distribution of information, Zhao *et al* (2000) examine conventional mailing lists and use the result of that examination to propose a new workflow mechanism. The design consists of the proposal of two new information distribution methods and an extension to existing information filtering algorithms. The paper does not develop a technology artefact as such, but proposes one (or more).

Characterisation using FEDS: A Technical Risk & Efficacy strategy were clearly used in this case. The artefact/evaluand developed was not instantiated, but was instead designed for two methods and an algorithm (also a kind of method). The evaluation is therefore clearly formative. The evaluation experiment was also artificial in that it used 'a very simple data set based … an example Seminar Announcement' (p. 67), an imaginary example created specifically for the purpose of the evaluation, which means the evaluation did not involve a real task (although it did involve real users). The properties chosen for evaluation were workflow possibilities (p. 70).

Analysis: The Technical Risk & Efficacy strategy reduced the technical risk by evaluating early, and was formative in that it would allow identification of problems with the algorithm to be fixed. It potentially would reduce costs by fixing technical difficulties before implementation or instantiation of the algorithms and workflows into a software system. The strategy used is also effective in controlling variables to demonstrate that the artefact(s) can deliver the improvement (in theory).

Suggestions for improvement: By itself, the evaluation strategy used may lack the potential for rigorous evaluation offered by summative evaluation strategies. However, summative evaluation may have been outside the scope of the research for various practical reasons. Furthermore, in the absence of other information, summative evaluation may already have been done at later stages of the DSR project, or indeed by other researchers.

A naturalistic summative evaluation of FEDS

The above examples illustrate FEDS by using it as a framework for understanding existing DSR work. In this section, we consider a DSR project that applied FEDS in deciding an evaluation strategy. We describe the actions taken for each of the four steps of the FEDS process and how the FEDS framework was used.

Kristiansen (2010) applied FEDS (although it did not have that name at the time) during his Ph.D. research to decide and design the evaluation strategy for the 'Site-Storming Method' artefact, which is aimed at development

of site-specific performative games that typically are used outdoors and not in front of a computer screen. The Site-Storming Method takes into account specifics of the 'site', using these specifics in the game. This *in situ* design method particularly contributed new ways of balancing creativity and structure in IS design. For example, the designed method accomplishes this by using IDEO-inspired cards (Best, 2006).

As described earlier, the first step in the FEDS process is to explicate the goals and constraints of the research. The goals for the Site-Storming method were that it should be applicable, usable, abstract, and formalised. Subject to the normal time constraints of Ph.D. research, Kristiansen had a limited time to be able to conclude his research, including both the design and the evaluation of the artefact and producing a thesis. As a DSR thesis, a reasonably strong level of rigour is expected and the artefact developed is also expected to work and make a useful contribution. Another issue that needed to be considered is that at the beginning of the research, there was a general lack of knowledge about site-specific games and their design, especially how humans would respond and react. This meant that, to some extent, the development would be exploratory in nature and it was likely that an initial design could have minor or even major flaws. Relating these to the design goals for the DSR evaluation, it would ultimately have to be rigorous, efficient enough to accommodate false starts, and formative in order to learn and improve the method as it developed. From an ethical point of view, there were few constraints except that any research participants should not be disadvantaged by participating. The developed artefact would not be safety critical, so technical rigour was not important from that perspective.

Step 2 in the FEDS process is to choose the evaluation strategy. Since formative evaluation was necessary and in order to conduct the research efficiently, with as few false starts and as little rework as possible, early formative evaluation of one or more designs was needed. However, the need for rigour in the Ph.D. thesis also indicated that a further summative evaluation would be needed.

Further, as speed and usability effectiveness were important, early artificial evaluation was indicated. However, once again the need for rigour indicated that naturalistic evaluation would also be needed. Hence using Table 1 it clearly would point to a Human Risk & Effectiveness Strategy with early artificial evaluation episodes followed by summative and naturalistic evaluation,

Step 3 in the FEDS process is to choose the properties to evaluate. In this case, the properties to be evaluated included properties of both the method itself as well as properties of the games developed by the method in order to evaluate the efficacy of the method. As noted above, properties of the method to be evaluated included its applicability and usability. Any resulting games should also be usable, fun, and encourage learning (these criteria based on theory of performative games).

The fourth step is to design the evaluation episodes. Ultimately, the research was conducted in three iterations. The first iteration resulted in the initial design of the Site-Storming Method. The second iteration formatively evaluated the initial design of the Site-Storming Method and developed a revised Site-Storming Method. The third iteration formatively evaluated the revised Site-Storming Method and developed a second revision of the Site-Storming Method. The Site-Storming Method design was validated using the framework presented in this paper 'by using the design process to design several game concepts' (Kristiansen, 2010, p. 37), where all the game concepts were 'evaluated *ex ante*, and one concept is implemented and evaluated *ex post*' using the game properties described above.

In the first iteration, the initial 'design' of the Site-Storming Method, the evaluation was conducted by applying a conceptual prototype of the approach to the development of a game called 'Gainers N' Drainers'. This evaluation was *formative* and *artificial* in that the application was by the researcher rather than real users, the artefact used was the design rather than the realised method, and the application was to the development of a hypothetical example game.

The second evaluation of the conceptual prototype of the method was to design and construct a real game for a 'larger project organization with several participants' (Kristiansen, 2010, p. 38). The game, called 'The Cliff Game' ('Klintespillet' in Danish), was developed for and with a real project owner, who wanted to engage children aged 12–15 in a physical exploration of the natural area around the famous chalk cliffs on the island of Møn, Denmark. This second evaluation of the initial design was again *formative*, because the Site-Storming Approach was still only conceptually designed. However, the evaluation was more *naturalistic* as it involved real users and a real problem. The evaluation resulted in important feedback for re-designing the Site-Storming Method.

In the second iteration, the Site-Storming Method was evaluated in three ways. First, it was evaluated by developing (conceptualising and describing) a total of 26 games during 3 different design sessions using the new method, in order to gather experiences and better understand the method. This was done *formatively* and *artificially* (not real users, games not intended for further development). Later in the second iteration a comparative study was conducted by developing two games 'The Ball' and 'Switch'. This evaluation was also *formative*, but more *naturalistic* since the games were both implemented.

In the third iteration, a revised, refined, and substantially completed version of the Site-Storming Method was evaluated in two different ways. First, it was evaluated by redesigning 'The Cliff Game' initially designed in the first iteration. Second, it was evaluated by applying the method in a game design workshop in which students worked with pervasive game design. Both evaluations were *summative*, in that the Site-Storming Method was complete and instantiated for use. The evaluation of the same game allowed comparison and was essentially *naturalistic* as 'The Cliff' Game was intended to be deployed, was

developed for real clients, and was tested with potential users. The second evaluation with students was less *naturalistic* as the designed games were not intended to be implemented or for real use (although they potentially could be).

The most important lesson from the Human Risk & Effectiveness strategy used by Kristiansen is that he responded to the general lack of knowledge about site-specific games and their design. He made extensive, iterative use of *formative* evaluation to evaluate different versions of the Site-Storming Method and to redesign and evolve the method as knowledge was gained about its requirements and issues with its design.

Another interesting aspect of this instance of using our evaluation framework was that different criteria were chosen for formative and for summative evaluations. As Kristiansen (2010) describes it for the formative evaluations: 'For the … evaluation the chosen criteria was [sic] efficacy and effectiveness, e.g. that the design process delivered several designs that were good examples …' (p. 38). And for the summative evaluations '… the criteria was [sic] efficacy and elegance, e.g. that the designed game lived up to expectations and that the design is aesthetically pleasing' (Kristiansen, 2010, p. 39). In both cases Kristiansen decided to choose criteria among the five E's (Checkland & Scholes, 1990).

This project illustrates how the strategic use of different kinds of evaluation episodes helps establish the quality of the knowledge delivered by the design science. The progress from formative and artificial evaluations towards more summative and more naturalistic evaluations adds rigour. This rigour strengthens not only the evidence about the utility of the Site-Storming Method in its real use environment, but also the concomitant knowledge contribution pertaining to the balance of creativity and structure. Concerning use of the framework, Kristiansen (2010) concludes that 'applying scientific methods to the evaluation is seen as necessary to recognize the design process as design science research' (p. 41).

In summary, the use of FEDS by Kristiansen to successfully determine an appropriate strategy for his DSR research provides useful evidence of the utility of the FEDS framework and process to determine and execute an appropriate evaluation strategy to successfully complete a DSR project. The case study further demonstrates the flexibility of the approach to accommodate difficult and conflicting goals and constraints in the context of a DSR project.

Discussion

There is a wide range of literature within and outside DSR that identifies different methods and paradigms for how to evaluate (e.g., in Nunamaker *et al*, 1990/1991; March & Smith, 1995; Hevner *et al*, 2004; Vaishnavi & Kuechler, 2004; Venable, 2006; Peffers *et al*, 2008; Sein *et al*, 2011; Kuechler & Vaishnavi, 2012). However, the prior literature provides little guidance on how (and why) to select appropriate methods or develop a strategy for what to evaluate, when, and how to conduct evaluation activities in DSR. This paper extends that literature by providing the FEDS framework and process as a means to guide DSR researchers towards the development of a suitable evaluation strategy to match a specific DSR project's situation.

FEDS is a novel evaluation framework uniquely suited to use in DSR. An important feature of the FEDS framework is its focus on the two key purposes of evaluation in DSR. In DSR, evaluation regards not only the utility aspect of the artefact in the environment, but also the quality of the knowledge contributed by the construction of the artefact. In order to better achieve both purposes, researchers can design a FEDS evaluation strategy that not only fits the goals and setting of the DSR project, but also sustains both the utility of the artefact in its environment and the transfer of knowledge to others.

Previous work largely assumes that DSR evaluations are monomorphic (only one kind of evaluation episode is needed). For example, Gill & Hevner (2013) provide guidelines for expanding evaluation to address not only utility, but also fitness and usefulness. But typically such guidance has not extended to dynamic evaluation designs that entail more than one kind of evaluation. Because design researchers can carry out more than one evaluation episode, more than one approach is possible in a single DSR project or programme. In fact many can be applied, as demonstrated in the previous section. It is possible to mix artificial and naturalistic evaluation as well as non-empirical, positivist, interpretive, and critical evaluation methods, supporting a pluralist view of science, where each has its strengths in contributing to a robust evaluation depending on the circumstance.

The validity and strength of an evaluation study for DSR is situated in the paradigm of the evaluation and how it evaluates the artefact's achievement of its intended purpose(s). An artificial evaluation derives strength in its validity from the reduction of reality to abstract properties, including those of the artefact and its intended surroundings and purpose. Such evaluations benefit from control and testing of these abstract properties and more directly link any results with efficacy in achieving the artefact's purpose. A naturalistic evaluation derives strength in its validity from actual performance against purpose in its intended environment. Abstraction and control are sacrificed for an examination of the artefact's effectiveness in achieving fulfilment of its purpose in the natural world. There are advantages to both artificial evaluation (such as more control, lower cost, and better theoretical validity) and naturalistic evaluation (such as more realism and better objective validity). Evaluation of artefacts in artificial settings is not limited to simple experimental settings, but includes somewhat imaginary or simulated settings where the technology (or its representation) can be studied under substantially artificial conditions.

Conclusion

How to go about choosing and designing an appropriate evaluation strategy (or approach) is a very significant issue

in IS DSR, which is under-addressed in the extant DSR literature. In this paper, we have developed and presented the FEDS framework and four-step evaluation design process based on an analysis and synthesis of works on evaluation in DSR and more generally across other domains.

We have also illustrated and evaluated the framework by using it to analyse two pre-existing DSR studies. We have further provided evidence of the effectiveness of the FEDS framework and process by describing a naturalistic summative evaluation study that applied the framework for developing a DSR evaluation strategy. These examinations highlighted features of the framework and provided evidence of its utility.

The contribution of FEDS is its unique suitability for guiding the design of effective strategies for the evaluation of design artefacts and design theories within DSR projects or programs. The evidence provided indicates that the FEDS framework and evaluation design process should help future DSR researchers (especially novice DSR researchers) to design and improve their DSR evaluation activities. The framework aids DSR researchers by offering a strategic view of DSR evaluation according to two dimensions: functional purpose (formative *vs* summative evaluation) and evaluation paradigm (naturalistic *vs* artificial evaluation). The formative perspective captures the possibility to reduce risk by evaluating early, before undergoing the cost and effort (possibly wasted) of building and rigorously evaluating an instantiation of a significantly flawed design for the artefact. The summative perspective offers the possibility of evaluating the instantiated artefact in reality, not just in theory or hypothetically. Naturalistic evaluation methods offer the possibility to evaluate the real artefact in use by real users solving real problems, while artificial evaluation methods offer the possibility to control potential confounding variables more carefully and prove or disprove design hypotheses, design theories, and the utility of design artefacts. The four different evaluation strategies identified should inspire DSR researchers to consider alternative evaluation strategies and the criteria identified should guide the appropriate choice of strategy. The four-step evaluation design process provides further guidance to DSR researchers on applying the FEDS framework and deciding what particular evaluation strategy(ies) to use on a particular DSR project/programme, by focusing their attention on different relevant aspects, such as goals and properties, in addition to the evaluation's functional purpose and paradigm inherent in the two dimensions of the framework.

While not a primary purpose of developing FEDS, the application of FEDS to two example DSR studies also shows that FEDS is also helpful in understanding the evaluation strategies of past DSR studies.

The research leading to and evaluating the FEDS Framework and Evaluation Design Process for DSR has limitations. The evaluation of the framework and process is limited to a small number of studies. Further application and evaluation, particularly on a variety of developed artefacts, would provide further validation of the approach. Another avenue for future work is to develop other novel strategies, and to explore further the value of hybrid strategies. The FEDS framework and evaluation design process would also benefit from further research to enhance their features as well as developing training materials to convey them more clearly to their future users.

Acknowledgements

This research was supported by the Curtin University School of Information Systems and the Curtin Business School.

About the authors

John Venable is Associate Professor and Director of Research in the School of Information Systems and Co-Director of the Not-for-Profit Research Initiative at Curtin University, Perth, Western Australia. His research interests include IS modelling and development methods, research methods, problem solving methods, knowledge management, and organisational culture and change management.

Jan Pries Heje is Professor in IS, Roskilde University, Denmark. He serves as Chair to IFIP Technical Committee 8 on IS. His research focuses on designing and building innovative solutions to managerial and organizational IT problems. Jan has published more than 200 books and papers in quality journals and conferences.

Richard L. Baskerville is a Board of Advisors Professor of Information Systems at Georgia State University and Professor in the School of Information Systems at Curtin University, Perth, Australia. His research and authored works regard security of information systems, methods of information systems design and development, and the interaction of information systems and organizations. Baskerville is Editor Emeritus and past Editor-in-Chief of the *European Journal of Information Systems*. He is a Chartered Engineer, holds a B.S. summa cum laude, from The University of Maryland, and the M.Sc. and Ph.D. degrees from The London School of Economics, University of London.

References

ALBERT TC, GOES PB and GUPTA A (2004) GIST: a model for design and management of content and interactivity of customer-centric web sites. *MIS Quarterly* **28(2)**, 161–182.

BASKERVILLE R, PRIES-HEJE J and VENABLE JR (2007) Soft design science research: extending the boundaries of evaluation in design science research. *2nd International Conference on Design Science Research in*

Information Systems & Technology (DESRIST 2007) (CHATTERJEE S and ROSSI M Eds), Claremont Graduate University, Pasadena.
BEST K (2006) *Design Management. Managing Design Strategy, Process and Implementation*. AVA Publishing SA, Lausanne, Switzerland.
CHECKLAND P and SCHOLES J (1990) *Soft systems Methodology in Practice*. J. Wiley, Chichester.
DELONE WH and MCLEAN ER (1992) Information systems success: the quest for the dependent variable. *Information Systems Research* **3**(**1**), 60–95.
GARVIN D (1987) Competing on the eight dimensions of quality. *Harvard Business Review* **65**(**6**), 101–109.
GILL TG and HEVNER AR (2013) A fitness-utility model for design science research. *ACM Transactions on Management Information Systems (TMIS)* **4**(**2**), 5-1–5-24.
GREGOR S and JONES D (2007) The anatomy of a design theory. *Journal of the Association for Information Systems* **8**(**5**), 312–335.
GUMMESSON E (1988) *Qualitative Methods in Management Research*. Studentlitterature Chartwell-Bratt, Lund.
HEVNER A, MARCH ST, PARK J and RAM S (2004) Design science in information systems research. *MIS Quarterly* **28**(**1**), 75–105.
IRANI Z and LOVE PED (2002) Developing a frame of reference for ex-ante IT/IS investment evaluation. *European Journal of Information Systems* **11**(**1**), 74–82.
KLECUN E and CORNFORD T (2005) A critical approach to evaluation. *European Journal of Information Systems* **14**(**3**), 229–243.
KRISTIANSEN E (2010) Computer games for the real world. Ph.D. dissertation. Roskilde University, Roskilde, Denmark.
KUECHLER W and VAISHNAVI V (2012) A framework for theory development in design science research: multiple perspectives. *Journal of the Association for Information Systems* **13**(**6**), 395–423.
MARCH ST and SMITH GF (1995) Design and natural science research on information technology. *Decision Support Systems* **15**(**4**), 251–266.
MATHIASSEN L, MUNK-MADSEN A, NIELSEN PA and STAGE J (2000) *Object Oriented Analysis & Design*. Marko Publishing, Aalborg, Denmark.
NUNAMAKER JF, CHEN M and PURDIN TDM (1990/1991) Systems development in information systems research. *Journal of Management Information Systems* **7**(**3**), 89–106.
PEFFERS K, TUUNANEN T, ROTHENBERGER MA and CHATTERJEE S (2008) A design science research methodology for information systems research. *Journal of Management Information Systems* **24**(**3**), 45–77.
PRIES-HEJE J, BASKERVILLE R and VENABLE JR (2008) Strategies for Design Science Research Evaluation. *Proceedings of the 16th European Conference on Information Systems (ECIS 2008), Paper 87* (GOLDEN W, ACTON T, VAN DER HEIJDEN H, CONBY K and TUUNAINEN VK Eds), Association for Information Systems, Atlanta, GA, USA, http://aisel.aisnet.org/ecis2008/87.
REMENYI D (1999) *IT Investment: Making a Business Case*. Butterworth-Heinemann, Woburn, MA.
SEIN MK, HENFRIDSSON O, PURAO S, ROSSI M and LINDGREN R (2011) Action design research. *MIS Quarterly* **35**(**1**), 37–56.
SMITHSON S and HIRSCHHEIM R (1998) Analysing information systems evaluation: another look at an old problem. *European Journal of Information Systems* **7**(**3**), 158–174.
STEFANOU CJ (2001) A framework for the ex-ante evaluation of ERP software. *European Journal of Information Systems* **10**(**4**), 204–215.
STOCKDALE R and STANDING C (2006) An interpretive approach to evaluating information systems: a content, context, process framework. *European Journal of Operational Research* **173**(**3**), 1090–1102.
STUFFLEBEAM DL (2003) The CIPP model for evaluation. In *International Handbook of Educational Evaluation* (KELLAGHAN T and STUFFLEBEAM DL, Eds), pp. 31–62, Kluwer Academic Publishers, Dordrecht, The Netherlands.
SUN Y and KANTOR PB (2006) Cross-evaluation: a new model for information system evaluation. *Journal of the American Society for Information Science and Technology* **57**(**5**), 614–628.
SYMONS VJ (1991) A review of information systems evaluation: content, context and process. *European Journal of Information Systems* **1**(**3**), 205–212.
VAISHNAVI V and KUECHLER W (2004) Design science research in information systems. AISNet. Association for Information Systems, Atlanta, Georgia, USA. January 20, 2004, last updated October 23, 2013, http://desrist.org/desrist/content/design-science-research-in-information-systems.pdf.
VENABLE R (2006) A framework for design science research activities. In *Emerging Trends and Challenges in Information Technology Management: Proceedings of the 2006 Information Resource Management Association International Conference* (KHOSROW-POUR M Ed), Idea Group Publishing, Hershey, PA, USA.
VENABLE JR, PRIES-HEJE R and BASKERVILLE R (2012) A Comprehensive Framework for Evaluation in Design Science Research. *7th International Conference on Design Science Research in Information Systems & Technology (DESRIST 2012)* (PEFFERS K and ROTHENBERGER MA, Eds), Springer, Berlin, Germany.
WALLS JG, WIDMEYER GR and EL SAWY OA (1992) Building an information system design theory for vigilant EIS. *Information Systems Research* **3**(**1**), 36–59.
WILLIAM D and BLACK P (1996) Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal* **22**(**5**), 537–548.
ZHAO JL, KUMAR A and STOHR EA (2000) Workflow-centric information distribution through e-mail. *Journal of Management Information Systems* **17**(**3**), 45–72.