A Taxonomy of Evaluation Methods for Information Systems Artifacts

NICOLAS PRAT, ISABELLE COMYN-WATTIAU, AND JACKY AKOKA

NICOLAS PRAT (corresponding author; prat@essec.edu) is an associate professor of information systems at ESSEC Business School, France. He holds a Ph.D. in information systems from the University of Paris Dauphine and a Master of Science in management from ESSEC Business School. His research interests include information systems design, business intelligence and analytics, knowledge management, and data provenance. His work has appeared in journals such as *Decision Support Systems, Data and Knowledge Engineering, Expert Systems with Applications*, and others, and in the proceedings of major academic conferences, including the International Conference on Conceptual Modeling (ER), Hawaii International Conference on Systems Sciences (HICSS), the Workshop on Information Technologies and Systems (WITS), and the European Conference on Information Systems (ECIS).

ISABELLE COMYN-WATTIAU is a professor of information and computer systems at the Conservatoire National des Arts et Métiers in Paris (CEDRIC-CNAM Research Center) and at ESSEC Business School. She received an M.S. and a Ph.D. in computer science from the University of Paris 6. Her research interests include information systems design, data warehouse design and redesign, and database integration. She has published more than seventy journal and conference papers on information and database systems. Her research has appeared in journals such as *Decision Support Systems, Data and Knowledge Engineering, Expert Systems with Applications*, and major academic conferences including the International Conference on Conceptual Modeling (ER) and the European Conference on Information Systems (ECIS).

JACKY AKOKA is a professor and holds the Chair in Information Systems at the Conservatoire National des Arts et Métiers in Paris (CEDRIC-CNAM). He received an M.S. in computer science and a doctoral degree in computer science and operations research from the University of Paris 6. He also received a Ph.D. in MIS from MIT Sloan School of Management. He is a professor of information systems at Télécom

A preliminary version of this paper was presented at the PACIS 2014 conference. The authors express gratitude to Mikko Siponen and Doug Vogel for their insightful suggestions. They also thank the three anonymous reviewers as well as the editor-in-chief for their generous input and helpful comments. This research benefited from the valuable help of Odette Sangupamba Mwilu in the content analysis work. The authors also acknowledge the suggestions of statistical colleagues for data analysis.

Journal of Management Information Systems / 2015, Vol. 32, No. 3, pp. 229–267.

ISSN 0742–1222 (print) / ISSN 1557–928X (online)
DOI: 10.1080/07421222.2015.1099390

École de Management (Institut Mines Télécom). He has published more than 120 conference and journal papers on information and decision systems. His research interests include information systems methodologies, decision support systems, and data warehouse design and implementation. His research has appeared in journals such as *Decision Support Systems, IEEE Transactions on Computers, Information and Management, Data and Knowledge Engineering*, and others, and in the proceedings of major academic conferences including the International Conference on Conceptual Modeling (ER) and the European Conference on Information Systems (ECIS).

ABSTRACT: Artifacts, such as software systems, pervade organizations and society. In the field of information systems (IS) they form the core of research. The evaluation of IS artifacts thus represents a major issue. Although IS research paradigms are increasingly intertwined, building and evaluating artifacts has traditionally been the purview of design science research (DSR). DSR in IS has not reached maturity yet. This is particularly true of artifact evaluation. This paper investigates the "what" and the "how" of IS artifact evaluation: what are the objects and criteria of evaluation, the methods for evaluating the criteria, and the relationships between the "what" and the "how" of evaluation? To answer these questions, we develop a taxonomy of evaluation methods for IS artifacts. With this taxonomy, we analyze IS artifact evaluation practice, as reflected by ten years of DSR publications in the basket of journals of the Association for Information Systems (AIS). This research brings to light important relationships between the dimensions of IS artifact evaluation, and identifies seven typical evaluation patterns: demonstration; simulation- and metric-based benchmarking of artifacts; practice-based evaluation of effectiveness; simulation- and metric-based absolute evaluation of artifacts; practice-based evaluation of usefulness or ease of use; laboratory, student-based evaluation of usefulness; and algorithmic complexity analysis. This study also reveals a focus of artifact evaluation practice on a few criteria. Beyond immediate usefulness, IS researchers are urged to investigate ways of evaluating the long-term organizational impact and the societal impact of artifacts.

KEY WORDS AND PHRASES: artifact evaluation, content analysis, design evaluation, design science, system design, taxonomy.

Artifacts should be at the core of information systems (IS) research [42]. In this context, IS artifact evaluation constitutes a major issue. Resorting to multiple research paradigms, IS artifact evaluation has traditionally been the responsibility of design science research (DSR). DSR builds and evaluates novel artifacts. In IS, DSR started with the paper of Nunamaker et al. [39], who proposed a multimethodological approach to IS research, centered on systems development. Even if they did not call them artifacts, the authors argued that data models or methods, for example, could contribute to theory building. DSR in IS gained wide acceptance after the paper by Hevner et al. [26]. It has a critical role to play in addressing major organizational and societal issues such as security [1, 54]. However, one has to admit that this research paradigm is still maturing—its theoretical bases are not yet

stabilized [16]. This is particularly true of artifact evaluation: DSR lacks commonly accepted, specific evaluation guidelines for the different artifact types [68].

In this paper, we investigate the "what" and the "how" of IS artifact evaluation: What are the objects of evaluation (evaluands), and the criteria against which to evaluate them? How—that is, with what methods—is evaluation carried out? What are the relationships between the different dimensions in the "what" and the "how" of evaluation? To answer these questions, we use the DSR paradigm to build and evaluate a taxonomy of evaluation methods for IS artifacts. The taxonomy comprises six dimensions: criterion, evaluation technique, form of evaluation, secondary participants, level of evaluation, and relativeness of evaluation. To structure the first dimension, we use the general systems theory: we consider evaluands as systems of artifacts, and organize the hierarchy of evaluation criteria according to the fundamental characteristics of systems. We apply the taxonomy of evaluation methods to analyze the content of 121 DSR papers from the basket of eight journals of the Association for Information Systems (AIS). The resulting database, composed of 402 evaluation methods, provides a unique picture of IS artifact evaluation practice, as reflected in ten years of publications in the AIS basket. The quantitative analysis of this database improves understanding of artifact evaluation in IS. This analysis highlights underexplored areas, where researchers should develop new evaluation methods, calling upon multiple paradigms (e.g., quantitative research, qualitative research, or action research). More specifically, methods are needed to evaluate the long-term organizational impact of IS artifacts and their societal impact, including ethicality and side effects. This study also illuminates relationships between the different dimensions of evaluation and leads to the identification of seven typical patterns in IS artifact evaluation.

Relevant Literature and Research Questions

We should differentiate design science (reflection and guidance on artifact construction and evaluation) from design research (construction and evaluation of specific artifacts) [68]. This work is based on a literature review of design science (focusing on artifact evaluation), combined with an analysis of a sample of design research papers. We summarize the design science literature and describe how we established the sample of design research papers. Finally, we formulate our research questions to fill the research gaps.

Review of the Design Science Literature

DSR in IS builds and evaluates artifacts, which may be constructs, models, methods, or instantiations [34]. The first IS paper relating to DSR is that of Nunamaker et al. [39], who propose a multimethodological approach to IS research, centered on systems development. Hevner et al. [26] propose a framework and guidelines for

rigorous and relevant DSR. To complete these guidelines, the design science research methodology of Peffers et al. [44] specifies the DSR process.

Beyond building and evaluating IS artifacts, some authors argue that DSR should develop theories. The concept of design theory in IS was originally developed by Walls et al. [60]. A design theory prescribes both the product (the properties an artifact should possess) and the process (the method for building the artifact, often referred to as design principles in other publications). The properties of the artifact should stem from the requirements, which are governed by so-called kernel theories (theories from the behavioral sciences). Testable product hypotheses evaluate whether the artifact meets its requirements (e.g., feasibility). Gregor and Jones [22] add two components to design theories: constructs and expository instantiation. The status, purpose, and content of design theories in IS remains heavily problematic [56]. In this paper, while acknowledging design theories as a possible output of DSR, we stick to the original typology of IS artifacts by March and Smith [34]. Considering the controversies mentioned above, adding theories to this typology is questionable. Moreover, this typology already enables the representation of typical components of design theories (e.g., constructs, expository instantiations, requirements as specific cases of models, and design principles as a specific case of methodologies).

Evaluation is critical in DSR. Even if the DSR process contains a large number of "micro-evaluations" [55], it is often divided into the build-and-evaluate activities at the macro level [26, 34, 68]. Peffers et al. [44] differentiate the activities of demonstration and evaluation. Demonstration uses the artifact to solve one or more problem instances, to establish that the artifact works. It is followed by more formal evaluation. The demonstration activity corresponds to the expository instantiation in the anatomy of a design theory [22]. Evaluation should demonstrate the utility of the artifact [26]. Some authors use the term "usefulness" as an apparent synonym for "utility" [44]. Beyond utility or usefulness, several criteria for artifact evaluation appear in the design science literature. The set of criteria suggested by March and Smith [34] is fragmented (being organized by artifact type) and most criteria are undefined. Other criteria are proposed [26, 50, 57], but the lists are incomplete, or the meaning of the criteria is unclear. To assess IS artifacts, several evaluation techniques (e.g., case study, field study, static analysis, and simulation) are mentioned in the literature [26, 39, 43, 47]. Several papers mix evaluation techniques (e.g., laboratory experiments) with form of evaluation (e.g., metrics) [47, 50, 57]. The variety of evaluation methods applied in DSR illustrates the influence of other research paradigms [28]. Some papers relate the "what" of evaluation (e.g., artifact types or evaluation criteria) with the "how" of evaluation (evaluation methods). Venable et al. [57, 58] present a framework for IS artifact evaluation, comprising two dimensions: naturalistic vs. artificial, and ex ante vs. ex post. Based on the "three realities" [52], naturalistic evaluation is characterized by real users using real systems (artifacts) to solve real problems (real tasks in real settings). Ex ante evaluation is formative—it takes place during artifact construction. Ex post evaluation is summative—it follows artifact construction. Within this

framework, an evaluation strategy is a planned trajectory along the two dimensions of evaluation. The authors propose a process to support the choice among evaluation strategies. Although this framework encompasses several important dimensions of evaluation in DSR, it remains at the macro level, and does not explicitly relate criteria with evaluation methods. Cleven et al. [11] propose a taxonomy of IS artifact evaluation. Evaluation criteria are absent from the taxonomy. Peffers et al. [43] relate types of artifacts with evaluation techniques. Sonnenberg and vom Brocke [50] list some evaluation techniques relevant for some evaluation criteria, but do not directly relate evaluation criteria with techniques.

This literature review illustrates that DSR is still maturing. The lists of evaluation criteria proposed in the literature are fragmented, often incomplete, and the meaning of the criteria is often unclear. The different dimensions of evaluation methods (the "how" of evaluation) are neither clearly distinguished, nor completely specified. The relationship between the "what" and the "how" remains unclear. A particular issue is the choice of evaluation methods according to criteria. Due to the lack of a conceptual definition of the different dimensions of evaluation, empirical analysis of artifact evaluation practice is difficult.

The next section describes the selection of the design research papers used in this research to analyze IS artifact evaluation practice, and the composition of the resulting sample.

Selection of Design Research Papers

To study artifact evaluation practice in IS, we coded and analyzed the content of design research papers. We chose a period of ten years, following the publication of Hevner et al. [26]. We considered the papers published in the AIS Senior Scholars' basket of journals: *EJIS, ISJ, ISR, JIT, JMIS, JSIS, JAIS*, and *MISQ*. This basket has often been used for citation or content analysis of IS research. Even though high-quality design research journals are omitted from the basket [9], it is a good source for high-quality DSR in IS.

This study is the first in-depth and longitudinal analysis of DSR evaluation practice based on the AIS basket of journals. We selected all papers from the basket that complied with the criteria described below. Compared with the present work, previous analyses of IS research based on the AIS basket were silent about DSR [31], limited to citation analysis [16], or focused on a specific research topic [4].

To select the papers, we limited ourselves to research articles and research notes, excluding editorials, opinion pieces, commentaries, review articles, and papers published online ahead of print.

The paper selection process comprised three steps: (1) *systematic check of the table of contents* for each journal from April 2004 to March 2014, (2) *keyword search* on Google Scholar to make sure no paper had been omitted in step (1), and (3) *exclusion of irrelevant papers during paper coding*. One of the authors performed steps (1) and

(2). Step (3) was performed collectively (any coder could suggest excluding a paper, but this decision had to be approved by all coders of the paper).

In the *systematic check of the tables of contents*, each paper was systematically screened starting from the title, and continuing with the abstract and full text as needed. In this first step, special care was taken not to exclude potentially relevant papers (if a paper appeared irrelevant upon closer examination, it could be excluded in the third step). We used four complementary, partially overlapping criteria: three sufficient conditions for paper exclusion from the sample and one sufficient condition for paper inclusion. The sufficient condition for paper inclusion was the explicit mention of DSR as the main research paradigm. The first condition for paper exclusion was that the main research paradigm was not DSR. Qualitative and quantitative research papers were often easy to identify. We also excluded papers in IS economics, and those using action research as a central approach. The second condition for exclusion was that the main objective of the paper was descriptive or explanatory. This criterion proved useful for papers that contribute mathematical models, often with no explicit reference to their research paradigm. If the mathematical model aimed primarily at understanding, we excluded the paper from the sample. If it was a model that could be customized and used in other contexts (e.g., as a component of a decision support system), we selected the paper. The third exclusion criterion was the absence from the paper of an IS artifact as a central contribution. To ease IS artifact identification, we applied our typology [46], which draws on Offermann et al. [41] and details the artifact types of March and Smith [34] as follows: construct (language, metamodel, concept), model (system design, ontology, taxonomy, framework, architecture, requirement), method (methodology, guideline, algorithm, method fragment, metric), and instantiation (implemented system, example).

The second step of paper selection consisted of a *keyword search* on Google Scholar. For each journal, we searched for the expression "design science" OR "design research" anywhere in the article, over the period. This resulted in a sample of 129 papers. The first step had only omitted a couple of relevant papers.

In the third step, we *excluded some papers from the sample after paper coding had started*, if detailed examination revealed that they were out of our research's scope. When coding a paper, we again used the first two conditions for paper exclusion applied in the first step. Moreover, since the content analysis of a paper coded its contributed IS artifacts and evaluation methods, the inability of coders to identify at least one artifact type and one evaluation method was a sign that we should consider the paper for exclusion from the sample. After the third step, the sample size fell to 121.

Table 1 shows the distribution of our sample by journal and period. It illustrates an increase in design research publications in the AIS basket, especially over the last period. Three clusters emerge: *JMIS* (30 papers), *EJIS*/*ISR*/*JAIS*/*MISQ* (20–22 papers), and *ISJ*/*JIT*/*JSIS* (2–3 papers). Online Supplement A [accessed at the publisher's website] provides the complete list of papers by journal.

Table 1. Design Research Papers by Journal and Period

	Apr.04–Mar. 06	Apr. 06–Mar. 08	Apr.08 – Mar. 10	Apr. 10–Mar. 12	Apr. 12–Mar. 14	Total
European Journal of Information Systems	2	8	5	2	3	20
Information Systems Journal	1	0	0	1	0	2
Information Systems Research	4	4	2	3	9	22
Journal of Information Technology	0	0	1	2	0	3
Journal of Management Information Systems	7	6	6	2	9	30
Journal of Strategic Information Systems	0	0	1	1	1	3
Journal of the Association for Information Systems	1	8	3	6	3	21
MIS Quarterly	1	0	5	6	8	20
Total	16	26	23	23	33	121

Research Questions

To fill the research gaps revealed by the literature review, we address the following questions:

> *RQ1: What are the objects of IS artifact evaluation and the criteria against which to evaluate them?*
>
> *RQ2: What are the evaluation methods, that is, the different options for evaluating artifacts?*
>
> *RQ3: What can we learn from IS artifact evaluation practice, as reflected by published research? More specifically, what are the frequently assessed and the unexplored criteria? What are the relationships between the different dimensions of evaluation, for example, what evaluation methods apply to what criteria? What typical evaluation methods emerge from practice?*

To answer these questions, we develop a taxonomy of evaluation methods for IS artifacts, following the DSR paradigm. We apply the taxonomy to analyze the content of the sample of design research papers.

Research Method for Developing the Taxonomy

To build and evaluate the taxonomy (our artifact), we apply the methodology for taxonomy development by Nickerson et al. [38]. We evaluate the taxonomy formatively and summatively, according to the ending conditions defined by these authors.

The Intended Outcome: A Taxonomy of Evaluation Methods

A taxonomy is a set of dimensions [38]. Each dimension consists of a set of two or more characteristics, such that for each object, each dimension has one and exactly one characteristic. This simple definition only allows flat dimensions. We also need hierarchical dimensions, grouping the characteristics (nodes) into categories. The highest category (root) comprises all characteristics. The other categories are subsets of the root. Formally, a taxonomy T may be defined as:

$$T = \left\{Dim_i, i = 1 \ldots, n \middle| Dim_i = \left\{Cat_{ij}, j = 1 \ldots, k_i\right\}\right|$$

$$Cat_{i1} = \{Char_{im}, m = 1 \ldots, p_i; p_i \geq 2\} \wedge \forall j \geq 2, Cat_{ij} \subsetneq Cat_{i1}\}$$

By convention, the first category (Cat_{i1}) is the root. Its name is the name of the dimension Dim_i. For flat dimensions, $k_i = 1$. Our taxonomy of evaluation methods comprises six dimensions:

$T = \{$Criterion; Evaluation technique; Form of evaluation; Secondary participants; Level of evaluation; Relativeness of evaluation$\}$

Let us consider an evaluation method and illustrate its representation with this taxonomy. Peffers et al. [44] demonstrate the efficacy of their DSR methodology by applying it retroactively to four already published IS research projects. In this evaluation method, the values of the six dimensions are (in that order): *efficacy, illustrative scenario, analysis or logical reasoning, none, real example or examples*, and *absolute*.

The Research Process

Based on the methodology proposed by Nickerson et al. [38], our research process for developing the taxonomy of evaluation methods is structured as shown in Figure 1 and detailed below.

Choice of Meta-Characteristic, Ending Conditions, and Approach for Taxonomy Development

This work investigates the "what" and the "how" of evaluation, and the way they correlate. Consequently, for the taxonomy of IS artifact evaluation methods, the "what" and the "how" are the *meta-characteristic*.

There are three types of *ending conditions* [38]: ending conditions that are part of the definition of a taxonomy, objective ending conditions, and subjective ending conditions. In our case, the applicable ending conditions depend on the phase of taxonomy development. Table 2 shows the ending conditions applicable in our approach, and the phases of the taxonomy development process (as represented in Figure 1) where these conditions apply. Details on applicability of ending conditions within the different phases are provided below.

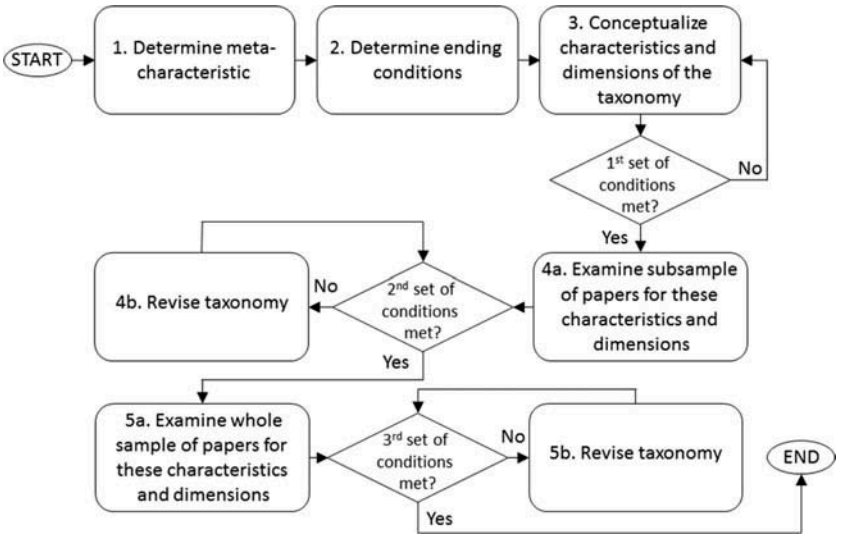

Figure 1. Methodology for Developing the Taxonomy of Evaluation Methods for IS Artifacts

Two *approaches to taxonomy development* are possible: conceptual to empirical and empirical to conceptual. Our approach is conceptual to empirical: we conceptualize an initial version of the taxonomy (phase 3), and use it to analyze and understand artifact evaluation practice through the examination of design research papers. This examination is based on content analysis [64]. Before examining our complete sample of papers (phase 5), we examine a subsample to test the coding scheme (phase 4). At each phase of taxonomy development, we test the applicable ending conditions and revise the taxonomy accordingly.

Conceptualization of the Characteristics and Dimensions of the Taxonomy

This phase uses the design science literature on artifact evaluation to conceptualize the taxonomy. In this conceptualization phase, most of the ending conditions in Table 2 apply. Objective ending conditions related to the examination or classification of objects are not applicable. We will consider the subjective ending conditions "robust" and "explanatory" once the hierarchy has been finalized (summative evaluation).

Examination of a Subsample of Papers and Revision of the Taxonomy

As recommended by the content-analysis literature, this phase tests the coding scheme on a subsample of papers, resulting in revisions to the taxonomy and coding heuristics (details are provided in the section "Content-Analysis Method").

Examination of the Whole Sample of Papers and Revision of the Taxonomy

This phase analyzes the content of the whole sample, resulting in a Microsoft Access database of 402 evaluation methods. This database is then analyzed quantitatively to understand the practice of IS artifact evaluation, with most of the advanced statistical analyses performed with R [45]. The results of this phase also provide a basis for further refining the taxonomy (more specifically, the hierarchy of criteria). After this phase, we evaluate the taxonomy summatively.

The ensuing sections detail the phases of taxonomy building and application (phases 3, 4, and 5 in Figure 1). We examine the applicable ending conditions in each phase. "Subjective" ending conditions [38] are indeed evaluation criteria, used for formative and summative evaluation of the taxonomy.

Conceptualization of the Characteristics and Dimensions of the Taxonomy

This phase (phase 3 in Figure 1) uses the IS design science literature on artifact evaluation. Based on the meta-characteristic, the taxonomy resulting from this conceptualization comprises six dimensions:

$$T = \{\text{Criterion; Evaluation technique; Form of evaluation; Secondary participants; Level of evaluation; Relativeness of evaluation}\}$$

The first dimension pertains to the "what" and the others detail the "how." An evaluation method is a unique combination of characteristics for the six dimensions. In the next sections, we discuss the "what" and the "how" of evaluation, and the verification of ending conditions for the conceptualized taxonomy.

The "What" of Evaluation

The "what" pertains to the objects of evaluation (aka evaluands) and to the criteria for evaluating these objects. We argue that the outputs of DSR (the evaluands) are systems of IS artifacts. This systems view forms the basis for organizing the evaluation criteria.

Design Science Research Outputs as Systems of Artifacts

Simon [48] describes artifacts in terms of their functioning and organization. He considers complex artifacts as hierarchies and stresses their internal operations and their interactions with the environment. Even though he points to some limitations of general systems theory (GST), his description of complex artifacts reflects a systems view. If complex artifacts are systems, IS artifacts should be no exception. This view is also supported in the IS literature. For example, Gregor and Iivari [21, p. 6] assert

Table 2. Ending Conditions Applied in the Development of the Taxonomy

Ending condition	Phases
Definition of a taxonomy	
The taxonomy consists of dimensions each with mutually exclusive characteristics.	3, 4, 5
The taxonomy consists of dimensions each with collectively exhaustive characteristics	3, 4, 5
Objective ending conditions	
All objects or a representative sample of objects have been examined.	5
At least one object is classified under every characteristic of every dimension.	5
Every dimension is unique and not repeated (i.e., there is no dimension duplication).	3
Every characteristic is unique within its dimension (i.e., there is no duplication within a dimension).	3, 4, 5
Subjective ending conditions	
Concise: limited number of dimensions and limited number of characteristics	3, 5
Robust: enough dimensions and characteristics to differentiate among objects	5
Comprehensive: all necessary dimensions and characteristics to classify objects of interest	3, 5
Extendible: easy inclusion of additional dimensions and characteristics	3, 4, 5
Explanatory: useful explanations of the nature of objects	5

that "Both an IS and an IT artifact qualify as a system because they have somewhere within their boundary a computer system that allows the artifact to change and exhibit mutability." The systems view also underlies the components of IS design theories [22], especially purpose and scope, principles of form and function, and artifact mutability.

Admittedly, some IS artifacts may not constitute systems in and of themselves. However, ultimately, a DSR endeavor should produce at least one instantiation and one abstract artifact (construct, model, or method), forming a system of artifacts. Among the typologies proposed in the GST literature, two are relevant here [2, 49]: the distinction between concrete and abstract systems, and between dynamic and static systems. Concrete systems have a physical existence. Abstract systems are made of concepts. In this sense, theories may be viewed as examples of abstract systems. The key elements of a theory are the concepts and the statements of relationships between them [19]. Dynamic systems possess structural components and activities. Static systems do not perform in themselves any kind of activity. GST often focuses on concrete, dynamic systems (e.g., living systems). Nonetheless, the systems view remains relevant for the other categories. For example, abstract systems can have an environment and undergo evolution, even if evolution and environment have different semantics for this category of systems [2, 49].

The typologies of systems mentioned above are related to the typology of artifacts [34]. Instantiations are concrete systems. Constructs, models, and methods are

abstract systems: they "do not have a physical existence, except in that they must be communicated in words, pictures, diagrams, or some other means of representation" [22, p. 321]. Constructs and models are static systems. Methods are dynamic.

Considering DSR outputs as systems of IS artifacts, we should evaluate these outputs based on the fundamental properties of systems. A system is a set of components in interrelation among themselves and with their environment [59, 65]. GST explains and describes the functioning and evolution of systems as a whole. It considers systems as holistic, open, goal-oriented, and self-organizing. A complex system is made of hierarchically organized subsystems. The subsystems communicate and work together to adapt to the environment. Several systems theories have been proposed [49]. Moreover, Bowler [7] characterizes a system as a hierarchy of subsystems sharing some common characteristics. Similarly, Churchman [10] emphasizes the role of users, decision makers, and designers of systems. All these approaches share the twelve fundamental properties of systems [49]: interrelationship and interdependence of objects, holism, goal seeking, transformation process, inputs, outputs, entropy, regulation, hierarchy, differentiation, equifinality, and multifinality. In order to reduce these fundamental properties to a more manageable set, we map them into the five aspects defining the canonical form of a system [30], respectively: goal, environment, structure, activity, and evolution. These five aspects encompass all the characteristics of a system upon which there is a consensus among the authors mentioned above. Table 3 illustrates the mapping between the fundamental properties of systems and the five aspects constituting the canonical form of a system (we omit entropy because this property only applies to living systems). Each of the eleven fundamental properties of systems maps into at least one aspect in the canonical form of a system. Thus, this canonical form suffices to structure the evaluation of DSR outputs. It will provide the basis for organizing the evaluation criteria.

Summing up, we view the objects of DSR evaluation (evaluands) as systems of IS artifacts. These artifacts may be concrete or abstract, and dynamic or static. Viewing DSR outputs as systems enables a holistic view of their evaluation: even though DSR produces several different artifacts, we should not consider these artifacts in isolation, and evaluation methods often assess the system as a whole, rather than a specific artifact. The systems view also provides the basis for organizing the hierarchy of criteria, along the five aspects in the canonical form of a system (goal, environment, structure, activity, and evolution). Depending on the nature of the artifacts comprising the DSR evaluand (concrete versus abstract and dynamic versus static), some criteria may not be relevant. For example, the criteria of the activity aspect apply only to dynamic systems. Nonetheless, considering DSR evaluands as systems provides an insightful perspective on their evaluation. Structuring the hierarchy of criteria along the five aspects of systems does not imply that we should evaluate DSR evaluands along all criteria of all aspects. The analysis of design research papers (section "Results from the Analysis of Design Research Papers") will reveal the most frequently assessed criteria and explore the relationships between these criteria and other dimensions of evaluation.

Table 3. Fundamental Properties vs. Canonical Form of Systems

	Goal	Environment	Structure	Activity	Evolution
Interrelationship and interdependence			X		
Holism			X		
Goal seeking	X				
Transformation process				X	
Inputs		X		X	
Outputs		X		X	
Regulation		X			X
Hierarchy			X		
Differentiation				X	
Equifinality	X				
Multifinality	X				

Sources: For fundamental properties, see [49]; for canonical form, see [30].

Hierarchy of Evaluation Criteria

We applied card sorting [36] to build the hierarchy of criteria. Benefits of this methodology include reduced subjectivity. It is relatively simple, allowing researchers to gain insight into the organization of information. It assumes that participants have knowledge of the information domain. The literature describes qualitative methods for analyzing the card sorting data, especially when the number of participants is low. Initially, we built a list of the most referenced papers on IS design science research. We ordered them by decreasing number of citations in Google Scholar. We extracted seventy-one evaluation criteria from these papers. We created one card (a Word document) for each criterion mentioned in each paper. It contained a name (the name denoting the criterion in the paper), a source (the reference of the paper), a definition (when provided), and comments (any complementary information on the criterion, taken from the paper).

The three authors performed card sorting. Systems theory allowed us to sketch categories to structure the set of evaluation criteria. Thus, in accordance with the five aspects of systems identified above, a closed sorting enabled the categorization of the seventy-one criteria in five sets: goal, environment, structure, activity, and evolution. Environment was split into three subcategories: people, organization, and technology [26]. For the first round, all researchers first performed an individual sorting. They agreed on a standard set of instructions. According to the latter, they had to sort each card into one or several categories. There is no correct predefined placement. Therefore, we could not use a hit ratio to measure the quality of the agreement between coders. We defined a metric as follows: if the participants did not agree on the categorization of a card, the measure was zero; otherwise, the measure was n if n participants agreed on the category. Finally, we compared this measure to the total agreement of the three coders on the seventy-

one criteria, leading to a ratio of 80 percent. No general authority exists with respect to required scores [62]. To improve the quality of our result, we conducted a brainstorming session in order to confront the disagreements and find the best categorization. The second round aimed at reducing the number of cards. Due to numerous identical names, we had to check whether criteria with identical or synonymous names could be merged (to comply with the ending condition that every characteristic in a taxonomy should be unique within its dimension); we also needed to detect potential homonyms. We performed an individual comparison of criteria located in the same categories thanks to the first round synthesis. Based on linguistic relationships, we proposed a merge when, reading the definitions, two or more criteria appeared to be synonyms or antonyms (e.g., simplicity and complexity). Finally, the third round chose the best name for synonyms and the positive term in case of antonyms (e.g., simplicity vs. complexity). If one term was the hypernym of another, we enriched the hierarchy with this link. We provide the final hierarchy in the Appendix, as part of the coding scheme used in the content analysis of the sample of papers. The Appendix shows the definitions for all criteria, citing the papers from which we took or adapted the definitions. The hierarchy of evaluation criteria constitutes the first dimension of our taxonomy. The specification of this dimension follows (characteristics are in italics):

Criterion = {Goal (Goal attainment (*Efficacy*; *Effectiveness*; *Validity*); *Utility*; Feasibility (...); *Generality*); Environment (...); Structure (...); Activity (...); Evolution (...)}

The "How" of Evaluation

The DSR literature distinguishes between naturalistic and artificial evaluation [57, 58]. In the former, real users use real artifacts to solve real problems (real tasks in real settings) [52]. Several dimensions operationalizing the "how" of evaluation reflect this distinction between naturalistic and artificial evaluation.

Evaluation Technique

The evaluation technique is a fundamental dimension [11, 26, 39, 47]. We define it as a hierarchical dimension, slightly adapting the typology of Hevner et al. [26]. Our category "question-based" includes survey [47] and focus group [53]. We also add action research. Considering that testing is marginal among evaluation methods [4], we do not decompose it into black-box and white-box testing.

Evaluation technique = {Observational or participatory (*Case study*; *Field study*; *Action research*); Analytical (*Static analysis*; *Dynamic analysis*); Experimental (Controlled experiment (*Laboratory experiment*; *Field experiment*); *Simulation*); *Testing*; Descriptive (*Informed argument*; *Illustrative scenario*); Question-based (*Survey*; *Focus group*)}

Form of Evaluation

As mentioned earlier, several design science papers mix the evaluation technique (e.g., laboratory experiment) and the form of evaluation (e.g., metric). To comply with the requirement that the characteristics within each dimension of a taxonomy should be mutually exclusive [38] (e.g., to express that a metric is used in the context of a laboratory experiment), the technique and the form of evaluation should be distinguished.

Cleven et al. [11] distinguish between quantitative and qualitative approaches to evaluation. We split quantitative into measured and perceived (aka latent) variables. Metrics are crucial in evaluation [34]. Perceived variables may be perceived directly or through items (latent-formative or latent-summative constructs). Other forms of evaluation are analysis [26] and logical reasoning [50], and formal proof [26].

Form of evaluation = {Quantitative (*Measured*; Perceived (*Perceived directly*; *Perceived through items*)); *Qualitative*; *Analysis and logical reasoning*; *Formal proof*}

Secondary Participants

Secondary participants take part in the evaluation of artifacts without being directly involved in their construction. They may be students, practitioners, or other researchers. This dimension relates to the distinction between naturalistic and artificial evaluation (real versus fake users). Students often provide less realistic conditions of evaluation than practitioners [66]. An evaluation method may imply several types of secondary participants. Consequently, we consider all possible combinations, to comply with the requirement that the characteristics within a dimension should be mutually exclusive.

Secondary participants = {*Students*; *Practitioners*; *Researchers*; *Students and practitioners*; *Researchers and practitioners*; *Students and researchers*; *Students*; *Researchers and practitioners*; *None*}

Level of Evaluation

This dimension corresponds to the distinction between ex ante evaluation (assessment of an abstract artifact) and ex post evaluation (assessment of an instantiated artifact) [57]. An instantiation may be an application of the artifact to a "real problem" [52] or to a fictitious one. We use the term "example" preferably to "problem" because an evaluated artifact (e.g., an algorithm) may be instantiated on a data set without explicit reference to a problem.

Level of evaluation = {*Abstract artifact*; Instantiation (*Fictitious example or examples*; *Real example or examples*)}

Relativeness of Evaluation

A new artifact should be better than existing artifacts [39]. Evaluation should thus establish its superiority by comparing it with extant solutions. The performance achieved with a new artifact may also be compared with the performance without the artifact.

Relativeness of evaluation = {*Absolute*; *Relative to absence of artifact*; *Relative to comparable artifacts*}

Ending Conditions and Formative Evaluation of the Taxonomy

The conceptualization of the characteristics and dimensions of the taxonomy was iterative. This was particularly the case for the hierarchy of evaluation criteria. In defining dimensions, we systematically checked that the characteristics of these dimensions were exclusive. For each dimension, we also watched out for missing characteristics to ensure exhaustiveness. Moreover, for the sake of exhaustiveness, the conceptualization of the taxonomy was based on all relevant design science papers dealing with IS artifact evaluation. At the end of the conceptualization of the characteristics and dimensions of the taxonomy, this taxonomy appears to comprise dimensions with mutually exclusive and collectively exhaustive characteristics. This will need confirmation in the empirical phases of taxonomy development (examination of design research papers). We should also underline that the notion of exhaustiveness is relative. We analyzed the relevant IS design science literature on artifact evaluation, but cannot definitely assert the exhaustiveness of this literature as regards the six dimensions of the taxonomy.

All six dimensions of the taxonomy are unique. This does not imply that they are orthogonal. For example, experimental evaluation methods (evaluation technique) operate on instantiations (level of evaluation).

In the conceptualization phase, three subjective ending conditions apply: the taxonomy should be concise, comprehensive, and extendible. In terms of our hierarchy of evaluation criteria, the taxonomy should be simple (structure/simplicity), complete (structure/completeness), and modifiable (evolution/modifiability). Modifiability is a hypernym of extendibility. We formatively assess these three criteria below.

To assess *simplicity*, we propose a specific metric (online Supplement B). According to this metric, the simplicity of the taxonomy resulting from the conceptualization phase is 0.21. The hierarchy of evaluation criteria is the most complex dimension of the taxonomy. The simplicity of this dimension, taken individually, is 0.26. We will compute these metrics again at the end of the taxonomy development process (summative evaluation).

At this stage, the *completeness* of the taxonomy of evaluation methods is assured by the systematic review of the IS design science literature on artifact evaluation. This is only one aspect of completeness. We will check this criterion again in the empirical phases of taxonomy development.

Finally, the hierarchical structure of the taxonomy makes it easy to *modify* by adding, deleting, or merging elements at a given level of this hierarchical organization.

Content-Analysis Method and Initial Test of the Coding Scheme

In the methodology for developing the taxonomy of evaluation methods, the empirical phases (phases 4 and 5 in Figure 1) apply the taxonomy to the content analysis of design research papers. Phase 4 tests the coding scheme on a subsample of papers, and revises the taxonomy and coding heuristics according to the results of this test. In this section, we detail phase 4 within the context of the content-analysis method followed. Phase 5 will be described in the following sections.

Content-Analysis Method

Coding Protocol

Content analysis bridges the gap between qualitative data and quantitative analyses [12]. It uses a coding scheme with coding units, categories, and coding rules. Coding assigns a category to a unit. The coding rules specify how to do this. Developing a coding scheme requires several steps [64]: (1) define the recording units, (2) specify the categories and the coding rules, (3) test the coding on a subsample of text, (4) assess reliability, (5) if the reliability is low, revise the coding rules and categories, and iterate to step (3) if necessary, (6) code the whole sample, and (7) assess reliability.

In this research, we have two levels of coding units. The first level is the journal article. For each paper in our sample, we coded the IS artifact or artifacts contributed by the paper, using the typology of artifacts [46]. The second level is the evaluation method. For each paper, we coded each method used by the authors to evaluate the artifacts, based on our taxonomy of artifact evaluation methods. Thus, the taxonomy was central to the coding scheme. The coding of an evaluation method determined the characteristic ("category," in content analysis vocabulary) for all six dimensions of the evaluation method. The Appendix shows an excerpt from the coding scheme. To ensure reliability, we also defined detailed coding heuristics and examples. This work required a series of workshops with all coders and resulted in a seventy-one-slide PowerPoint document.

Our recording units precluded using content analysis software. More specifically, the terms denoting the evaluation criteria (e.g., "performance") are often ambiguous, which required constantly going back to the definitions of criteria in the coding scheme. The three authors as well as a research assistant participated in the coding. The coding was performed with Microsoft Access. The individual Access databases were then used for automated computation of intercoder reliability, and the integrated database for data analysis was derived from these databases once intercoder differences were discussed and resolved.

We tested the coding scheme on a subsample of ten papers. Subsampling took into account the proportions of papers in the different journals and sought diversity in the contributed artifacts and evaluation methods. Each author and the research assistant coded the ten papers, and intercoder reliability was computed. Based on the results from the test, the taxonomy and the coding heuristics were revised. The modified

coding scheme was then applied to the whole sample. In this phase, all papers were double-coded. To ensure continuity in coding and in discussions over disagreements, one author coded all papers. After double coding, reliability was computed again. We detail reliability assessment and improvement below.

Reliability Assessment and Improvement

We considered using Cohen's kappa, a popular measure of intercoder reliability. However, this measure assumes that the different raters have coded the same elements (e.g., the same paragraphs of the same texts). In our case, the difficulty comes from the fact that evaluation methods (our recording units) are not clearly identified in DSR papers. Two coders may find different methods in the same paper, and the correspondence between these methods is not known a priori. Consequently, we had to develop specific metrics.

We defined two metrics of intercoder reliability, denoted as M_1 and M_2. These metrics are computed automatically for each paper, for each pair of coders (C_i, C_j). They are based on a one-to-one mapping between the evaluation methods identified by C_i and those identified by C_j. For each paper, the algorithm considers the coder who has identified the smallest number of evaluation methods (between C_i and C_j), and maps each evaluation method of this coder to an evaluation method of the other coder. Among the candidate mappings, the algorithm chooses the best match, that is, the one that minimizes the average distance between evaluation methods. The distance between two evaluation methods is the average distance between the characteristics of the two evaluation methods, for each dimension in the taxonomy of evaluation methods. The dimensions may be flat or hierarchical. For hierarchical dimensions, we apply the generalization distance [51]. Based on the mappings between evaluation methods, for each paper and each pair of coders (C_i, C_j), metric M_1 is the percentage of mapped evaluation methods. It measures the extent to which coders C_i and C_j have identified the same number of methods. Metric M_2 is the average distance between mapped evaluation methods. These two metrics are good substitutes for Cohen's kappa, given the peculiarities of the coding situation and the necessity of establishing a posteriori the correspondence between coded evaluation methods.

After coding of the subsample, the computed average percentage of mapped evaluation methods was 59 percent, with an average distance of 0.31 between mapped evaluation methods. These quite low scores required some changes in the coding scheme. To make these, we held a workshop in which all coders participated. Based on analysis of the differences in coding, the workshop discussed revisions of the taxonomy of evaluation methods, additions to the coding heuristics and examples, and refinements of the heuristics. As a result, the size of the document detailing and exemplifying the heuristics was doubled. After double coding of all papers with the improved coding scheme, the average percentage of mapped evaluation methods reached 71 percent, with an average distance of 0.21 between mapped evaluation methods. Considering the progress in intercoder reliability, the relative complexity of

the taxonomy used for coding, and the specificity of the coding context, we believe these values are reasonable. After computation of intercoder reliability, the coding disagreements were discussed and resolved between each pair of coders, leading to a database of 402 evaluation methods.

Examination of a Subsample of Papers and Revision of the Taxonomy

The examination of the subsample of ten papers (phase 4a in Figure 1) led to changes to the taxonomy (phase 4b). These changes followed from the verification of the ending conditions, or were suggested by recommendations from the content analysis literature. Actually, these recommendations partly overlap with the ending conditions of the methodology for taxonomy development. For example, content analysis authors suggest that categories should generally be mutually exclusive [64] ("categories" in content analysis vocabulary correspond to "characteristics" in the taxonomy vocabulary presented earlier).

The first change to the taxonomy decomposed the characteristic "testing" into black-box and white-box testing [26]. We did not make this distinction initially, due to the rarity of testing as an evaluation method. However, the analysis of disagreements between the coders of the subsample, as well as the workshop in which we discussed these disagreements, revealed the ambiguity of the term "testing." In content analysis, one of the chief objectives of trying the coding scheme is the resolution of ambiguities. The approach applied here (splitting a category into subcategories) is a classical disambiguation technique.

The second change concerned the dimension "relativeness of evaluation." Conceptually, "absolute," "relative to absence of artifact," and "relative to comparable artifacts" appeared as mutually exclusive, thus complying with the first ending condition in the definition of taxonomies. However, the content analysis of the subsample revealed difficulties in distinguishing between "absolute" and "relative to absence of artifact." We thus decided to keep only the characteristics "absolute" and "relative to comparable artifacts." "Absolute" meaning that the artifact is not compared to others, it encompasses "relative to absence of artifact."

Finally, inside the dimension "form of evaluation," the test of the coding scheme revealed the ambiguity of the term "analysis *and* logical reasoning," which seemed to suggest that these two forms of evaluation should always appear together. To comply with the ending condition that the characteristics within a dimension should be collectively exhaustive, we renamed this characteristic as "analysis *or* logical reasoning."

Results from the Analysis of Design Research Papers

This section describes the next phase of the taxonomy development methodology (phase 5a). In this phase, we analyzed the content of the sample of 121 design research papers. The data resulting from content analysis provided the basis for

quantitatively analyzing and understanding the practice of IS artifact evaluation. We present the important results below.

Slow Maturation of DSR Practice

Based on our typology of artifacts [46], we computed the frequencies by artifact type and subtype (the frequency of an artifact type or subtype is the percentage of papers contributing an artifact of this type or subtype). The most notable finding is the primacy of methods over models. Within methods, the most common subtypes are algorithm (35 percent), methodology (26 percent), and guideline (17 percent). Within models, all frequencies are below 15 percent. The primacy of methods over models is a sign that DSR practice is maturing.

We found 402 methods for artifact evaluation in the sample of papers (3.3 methods per paper on average). The average number of evaluation methods per paper by period has regularly increased since April 2006, reaching 3.9 over the last period.

To assess the diversity of evaluation methods in the different periods, we calculate the number of unique evaluation methods in each two-year period, that is, all the combinations of (criterion, evaluation technique, form of evaluation, secondary participants, level of evaluation, relativeness of evaluation) in the papers of this period. "Unique" means that for each period, we count the combinations appearing in several papers of the period only once. The diversity of evaluation methods constantly increases, from thirty-nine in the first period to eighty-nine in the last period. If we divide the number of unique evaluation methods by the number of papers in each period, this ratio decreases between the first and the second period (2.44 to 1.85), but it then increases regularly, reaching 2.70 in the last period. The increased diversity of evaluation methods is yet another sign that IS artifact evaluation practice is maturing. We complete this analysis by assessing the diversity of evaluation methods for the evaluation criteria. To this end, we compute the number of unique methods for each criterion, over the whole perimeter of the 121 papers. The criterion with the most diversity in evaluation methods is usefulness (thirty-five unique evaluation methods). A closer look at evaluation methods for this criterion reveals variety in evaluation techniques (case study, action research, laboratory experiment ...) and forms of evaluation (measured, perceived through items, qualitative ...). This variety illustrates the influence of diverse paradigms (including qualitative and quantitative research) on artifact evaluation in IS.

Table 4 shows the frequency of each form of evaluation (number and percentage of papers using each form of evaluation). Analysis or logical reasoning predominates (for example, in demonstrating the efficacy of artifacts with illustrative scenarios, as the cluster analysis below will confirm). Metrics appear in sixty-five papers (54 percent). Qualitative and quantitative research enrich the forms of evaluation. For example, usefulness may be assessed based on the scale of Davis [14]. Formal proof remains marginal in artifact evaluation.

An analysis by evaluation approach reveals that descriptive approaches still predominate, despite the call of Hevner et al. [26] to use them sparingly: these approaches are used in 52.9 percent of the papers and are followed by experimental approaches (48.8 percent), analytical approaches (19.8 percent), and observational or participatory approaches (18.2 percent). These results also confirm that in the "real versus lab debate," "Laboratory-based DSR remains prevalent" [33, p. 174]: empirical approaches dominate over observational or participatory approaches. A detailed analysis by evaluation technique shows the prevalence of simulation within experimental approaches (37 percent of the papers in our sample), and illustrative scenarios within descriptive approaches (48 percent). Case studies appear in only 13 percent of the papers. Per the definitions in our coding scheme, the term "case study" implies a real-world problem-solving situation. This term is often abused in the DSR literature, and a close examination of the papers revealed that many "case studies" were illustrative scenarios.

In summary, our data show that DSR, and more specifically artifact evaluation, is slowly maturing. While artifact evaluation practice is diversifying, some major tendencies remain (e.g., the prevalence of descriptive and experimental approaches).

Frequently Assessed and Unexplored Evaluation Criteria

Table 5 shows the frequency of evaluation criteria (number of papers assessing each criterion). The most common criteria are efficacy (N = 97), usefulness (N = 42), technical feasibility (N = 39), accuracy (N = 34), performance (N = 28), effectiveness (N = 22), ease of use, robustness, scalability, and operational feasibility (N = 12). Evaluation should establish that the artifact meets its goal and is useful. Thus, finding efficacy, usefulness, and effectiveness among the top criteria is not surprising. Like usefulness, ease of use is a common criterion in IS literature [14]. Technical and operational feasibility are the most frequently assessed categories of feasibility. Finally, accuracy, performance, robustness, and scalability are often evaluated with metrics. Readily available metrics may partly explain the high evaluation frequency of these criteria.

Among the thirty-nine criteria, seventeen (44 percent) are never evaluated in our sample. More specifically, a key finding from this study is that ethicality and side effects are never evaluated. Regarding ethicality, this may be explained by the absence of evaluation methods for assessing this criterion. As concerns side effects, they can only be fully assessed in the long term, which is not necessarily compatible with the time horizon of journal publications. It is time for the DSR community to develop approaches for fully evaluating the organizational and societal impact of artifacts. This includes approaches for evaluating ethicality [37] and side effects. Moreover, alignment with business is never assessed in the sample, and economic feasibility is only assessed once. This suggests partial disconnect between DSR and business-specific concepts and issues. Finally, many of the criteria pertaining to the structure of artifacts are never

Table 4. Frequency of Form of Evaluation by Journal

		EJIS		*ISJ*		*ISR*		*JIT*		*JMIS*		*JSIS*		*JAIS*		*MISQ*		Total	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Analysis or logical reasoning		20	100	2	100	16	73	3	100	23	77	3	100	21	100	18	90	106	88
Formal proof						8	36			3	10							11	9
Qualitative		4	20			2	9			5	17	1	33	4	19	7	35	23	19
Quantitative	Measured	4	20	1	50	16	73	2	67	24	80	1	33	6	29	11	55	65	54
	Perceived directly							1	33	4	13					3	15	8	7
	Perceived through items					1	5			9	30					2	10	12	10

assessed. DSR would benefit from adapting metrics from the software and data quality literature, for example, to assess the structural simplicity of artifacts.

Compositional Styles in IS Artifact Evaluation

To discover patterns in artifact evaluation, we clustered the database of 402 evaluation methods. Using R [45], we performed hierarchical clustering with average linkage, in-line with the suggestions of Hair et al. [24]. We preferred hierarchical clustering over *k*-means because we did not know the number of clusters in advance. We sought to discover the most typical compositional styles,[1] that is, the most typical patterns in artifact evaluation. Therefore, our interest was on the largest clusters. We did not need clusters of homogeneous sizes, hence our preference for average linkage over Ward's method.

Because the six dimensions of evaluation methods were symbolic, we needed a specific distance for hierarchical clustering. We reused the distance defined for reliability assessment (subsection "Reliability Assessment and Improvement" above). From the generated dendrogram, we chose the cluster solution so as to minimize increase in heterogeneity while maximizing the percentage of evaluation methods classified in clusters of size twenty or above (a size of twenty corresponds to about 5 percent of the total population of evaluation methods). Cutting the dendrogram at an agglomeration coefficient of 0.30 was the best compromise. This solution comprises seven clusters of size twenty or above, representing 78 percent of the population of evaluation methods. For all seven clusters, the average within-cluster distance is 0.2 or below. Table 6 shows the clusters, with their centroids. We define a cluster centroid as an evaluation method (i.e., a unique combination of characteristics for the six dimensions of the taxonomy) such that the average distance with the other evaluation methods of the cluster is minimal. We derived the names of the clusters from their centroids and the evaluation methods closest to the centroids. Each cluster represents a compositional style in IS artifact evaluation. The most common compositional style (largest cluster in Table 6) is *demonstration*. This style is typically used to demonstrate, by analysis or logical reasoning, that the artifact works (efficacy, technical feasibility), based on an illustrative scenario, using real or fictitious examples. It does not require secondary participants. The second most common compositional style is *simulation- and metric-based benchmarking of artifacts*: the efficacy, accuracy, performance, robustness or scalability of the artifact is measured and compared with those resulting from other approaches. Similarly to demonstration, this style does not require secondary participants. *Practice-based evaluation of effectiveness* typically establishes effectiveness of the artifact in a real setting. The evaluation techniques are observational or participatory, and practitioners take part in the evaluation. Like demonstration, the typical form of evaluation is analysis or logical reasoning. The next compositional style is *simulation- and metric-based absolute evaluation of artifacts*. This style is similar to the second style and assesses the same criteria. The difference lies in the relativeness of evaluation. In the style *practice-based evaluation of usefulness or ease of use*, the form

Table 5. Frequency of Assessed Criteria by Journal

	EJIS		*ISJ*		*ISR*		*JIT*		*JMIS*		*JSIS*		*JAIS*		*MISQ*		Total	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Goal/Goal attainment/Efficacy	15	75	1	50	20	91	3	100	27	90	2	67	15	71	14	70	97	80
Environment/People/Usefulness	8	40			5	23	1	33	9	30	1	33	6	29	12	60	42	35
Goal/Feasibility/Technical feasibility	7	35	1	50	5	23	2	67	8	27			7	33	9	45	39	32
Activity/Trustworthiness/Accuracy			1	50	5	23	1	33	14	47	1	33	5	24	7	35	34	28
Activity/Performance	1	5			9	41			10	33	1	33	4	19	3	15	28	23
Goal/Goal attainment/Effectiveness	5	25	1	50	2	9			2	7	1	33	5	24	6	30	22	18
Environment/People/Ease of use	2	10			1	5			3	10	1	33	2	10	3	15	12	10
Evolution/Robustness					5	23			6	20			1	5			12	10
Evolution/Scalability					2	9			6	20			3	14	1	5	12	10
Goal/Feasibility/Operational feasibility	3	15	1	50	2	9			1	3			2	10	3	15	12	10
Goal/Utility	2	10	1	50									4	19	1	5	8	7
Goal/Goal attainment/Validity									4	13			2	10	1	5	7	6
Structure/Completeness	1	5			2	9							3	14	1	5	7	6
Structure/Homomorphism/Fidelity to modeled phenomena	2	10							2	7			1	5			5	4
Activity/Completeness	2	10							2	7							4	3
Evolution/Adaptability	1	5											2	10			3	2
Activity/Trustworthiness/Reliability													1	5	1	5	2	2
Evolution/Learning capability															2	10	2	2
Activity/Simplicity									1	3							1	1
Goal/Feasibility/Economic feasibility	1	5															1	1
Goal/Generality									1	3							1	1
Structure/Homomorphism/Correspondence with another model/Construct deficit	1	5															1	1

of evaluation is typically qualitative. *Laboratory, student-based evaluation of usefulness* typically measures the task performance of students using the artifact or different variations of the artifact. Finally, *algorithmic complexity analysis* investigates the time or space complexity of the artifact (typically, an algorithm), by formal proof or by analysis or logical reasoning (this cluster comprises two centroids).

Relationships Between and Within the "What" and the "How" of Evaluation

To investigate the relationships between the "what" and the "how" of evaluation, we started by crossing the most frequently assessed evaluation criteria (the "what") with the other five dimensions of evaluation (the "how"). The analysis of the number of evaluation methods by criterion and evaluation technique shows that for efficacy, illustrative scenarios predominate. This illustrates again the commonality of demonstration as a compositional style. For usefulness, the three most common techniques are (in order): illustrative scenarios, laboratory experiments, and case studies. Accuracy, robustness, and scalability are generally assessed by simulation. Simulation is also the common evaluation technique for performance, in equal proportion with analytical approaches (analysis of complexity). The analysis by evaluation criterion and form of evaluation shows that for efficacy, analogical or logical reasoning predominates (it is used in 71 out of 110 methods evaluating efficacy, and the combination of efficacy with analysis or logical reasoning is typical of the demonstration compositional style). Metrics are used in 32 out of 110 methods evaluating efficacy. More specifically, papers contributing an optimization model may use the value of the objective function as a measure of this criterion. Mathematical models may also be analyzed formally to establish efficacy. Regarding usefulness, the most common forms of evaluation are qualitative evaluation and metrics. Metrics for usefulness often assess task performance of users. Accuracy is generally assessed by precision, recall, or combinations of these metrics. Finally, our data suggest that the assessed criterion influences the relativeness of evaluation. For example, efficacy evaluation is generally absolute. Form and relativeness of evaluation are related (most of the twenty-five methods that evaluate efficacy relatively are based on metrics). For accuracy, evaluation is predominantly relative (benchmarking of accuracy with comparable artifacts).

To explore more systematically the relationships between the different dimensions of evaluation, we performed in-depth statistical analyses using XLSTAT. We found no significant correlation between the six dimensions in our taxonomy of evaluation methods, thus confirming that all dimensions are clearly distinct. As mentioned above, this does not imply that the latter are orthogonal. To analyze more finely interdependencies between the dimensions of evaluation, we performed association rule mining with R [23]. The transactions were the 402 evaluation methods and the variables were the six dimensions of the taxonomy.

Table 6. Top Seven Clusters of Evaluation Methods, with Centroids

	Criterion	Evaluation technique	Form of evaluation	Participants	Level of evaluation	Relativeness	Size	Avg. dist.
Demonstration	Goal/Goal attainment/ Efficacy	Descriptive/Illustrative scenario	Analysis or logical reasoning	None	Instantiation/Real example or examples	Absolute	99	0.16
Simulation-, metric-based benchmarking of artifacts	Goal/Goal attainment/ Efficacy	Experimental/Simulation	Quantitative/Measured	None	Instantiation/Real example or examples	Relative to comparable artifacts	56	0.13
Practice-based evaluation of effectiveness	Goal/Goal attainment/ Effectiveness	Observational or participatory/Case study	Analysis or logical reasoning	Practitioners	Instantiation/Real example or examples	Absolute	45	0.18
Simulation-, metric-based absolute evaluation of artifacts	Activity/Performance	Experimental/Simulation	Quantitative/Measured	None	Instantiation/Real example or examples	Absolute	34	0.15
Practice-based evaluation of usefulness or ease of use	Environment/People/ Usefulness	Observational or participatory/Case study	Qualitative	Practitioners	Instantiation/Real example or examples	Absolute	34	0.18
Laboratory, student-based evaluation of usefulness	Environment/People/ Usefulness	Experimental/Controlled experiment/ Laboratory experiment	Quantitative/Measured	Students	Instantiation/Real example or examples	Absolute	24	0.20
Algorithmic complexity analysis	Activity/Performance	Analytical/Dynamic analysis	Formal proof	None	Abstract artifact	Absolute	22	0.16
	Activity/Performance	Analytical/Dynamic analysis	Analysis or logical reasoning	None	Abstract artifact	Absolute		

Table 7 illustrates some rules resulting from this analysis (rules R1 to R4). We may interpret these rules as follows: *Generally, the evaluation of the efficacy criterion with an illustrative scenario permits absolute evaluation of this criterion* (rule R1). *Generally, when evaluation is performed with practitioners, using real examples, this results in absolute evaluation* (rule R2, suggesting that this specific combination of secondary participants and level of evaluation may not be the most appropriate for comparative evaluation). *If no secondary participants are available, the evaluation technique of simulation may be considered* (rule R3). *If one wishes to evaluate artifacts relatively using metrics, simulation may be an appropriate evaluation technique* (rule R4). In discovering rules, we used the values of 0.8 and 0.1 for minimum confidence and minimum support, respectively. These thresholds are the defaults in the R apriori function, and are commonly used (e.g., [13]). With these constraints on confidence and support, we found more rules than those shown in Table 7, which illustrates some of the most relevant rules.

To complete this analysis, we performed association rule mining again, this time including the artifact types contributed by each article. Because we promote a systems view of IS artifact evaluation and do not relate directly artifact types with evaluation methods, the transactions considered for this analysis were the articles (thus, the number of transactions was reduced from 402 to 121, requiring more care in interpreting the results). The goal of the analysis was to relate the artifact types with the assessed criteria and the evaluation techniques. We considered each article as a "basket" composed of all artifact types, evaluation criteria, and evaluation techniques for the article. Table 7 illustrates two of the rules resulting from this analysis: *Using an illustrative scenario with an example is a possible way to evaluate the efficacy of a methodology* (R5); *Simulation may be recommended to evaluate the accuracy of an algorithm* (R6).

Evaluation

In this section, we evaluate the taxonomy of evaluation methods. This evaluation follows the examination of the sample of 121 papers (phase 5a in Figure 1), detailed in the previous section. We examine the ending conditions and formatively evaluate the taxonomy, leading to revisions (phase 5b). These revisions pertain to the hierarchy of criteria within the taxonomy. Finally, we summatively evaluate the resulting taxonomy.

Ending Conditions and Formative Evaluation of the Taxonomy

At this stage, we need to check that all objects or a representative sample of objects have been examined. In our case, the objects are the evaluation methods in the sample of 121 papers. The process for selecting these papers (explained in the subsection "Selection of Design Research Papers" above) ensures the representativeness of the sample. The requirement that at least one object should appear under every

Table 7. Relationships Between and Within the "What" and the "How" of Evaluation: Association Rule Mining

Rule #	Rule formal expression	Confidence	Support
R1	{Eval_technique=Descriptive/Illustrative scenario, Criterion=Goal/Goal attainment/Efficacy} => {Relativeness_of_eval=Absolute}	1.00	0.13
R2	{Secondary_particip=Practitioners, Level_of_eval=Instantiation/Real example or examples} => {Relativeness_of_eval=Absolute}	0.94	0.20
R3	{Eval_technique=Experimental/Simulation} => {Secondary_particip=None}	0.92	0.27
R4	{Form_of_eval=Quantitative/Measured, Relativeness_of_eval=Relative to comparable artifacts} => {Eval_technique=Experimental/Simulation}	0.87	0.15
R5	{Eval_technique=Descriptive/Illustrative scenario, Artifact_type=Instantiation/Example, Artifact_type=Method/Methodology} => {Criterion=Goal/Goal attainment/Efficacy}	1.00	0.11
R6	{Criterion=Activity/Trustworthiness/Accuracy, Artifact_type=Method/Algorithm} => {Eval_technique=Experimental/Simulation}	0.94	0.14

characteristic of every dimension is not met: as mentioned earlier, seventeen out of thirty-nine criteria in the hierarchy of criteria are never assessed in the sample. These criteria are listed below (we denote them with their categories in case of ambiguity): ethicality, environment/people/absence of side effects, alignment with business, environment/organization/absence of side effects, fit into technical IS architecture, alignment with IT innovation, environment/technology/absence of side effects, structure/simplicity, style, structure/consistency, construct overload, construct redundancy, construct excess, functionality, activity/consistency, efficiency, and modifiability. We will consider this list of criteria when simplifying the hierarchy.

In the conceptualization of the dimensions and characteristics of the taxonomy, we considered that the taxonomy was complete ("comprehensive," in terms of ending conditions) inasmuch as it was based on a review of the relevant design science literature. We should now verify that the dimensions of the taxonomy include all characteristics of objects, that is, of evaluation methods found in the sample of papers. More specifically, we should consider the need to include additional criteria in the hierarchy of criteria, in the case of criteria assessed in the sample of papers but absent from the hierarchy. In the coding scheme (Appendix), coders could suggest additional criteria. When the pairs of coders met to discuss and resolve coding disagreements, they also confronted and discussed additional criteria suggested by each coder. This provided the basis for deciding about revisions to the hierarchy of criteria to improve its completeness, as explained below.

Revision of the Hierarchy of Criteria

We used the suggestions for additional criteria and ensuing discussions to improve the completeness of the hierarchy. Regarding simplicity, we considered the seventeen criteria above as potential candidates for removal.

Improving Completeness

Among the criteria used in the sample but absent from our hierarchy, several criteria (e.g., reusability) were only applicable to specific categories of artifacts. We did not add these criteria to the hierarchy, in order to preserve its general applicability. Eventually, we added one criterion and extended the definition of another criterion. We added structure/understandability. This criterion appears in the design science literature [34], but we did not include it initially, considering it as a hyponym of people/ease of use. The examination of the sample revealed that for some artifacts, understandability is clearly distinct from ease of use, hence the necessity to add it to the hierarchy. Adapting a definition from software engineering [29], we define understandability as *the degree to which the artifact can be comprehended, both at a global level and at the detailed level of the elements and relationships inside the artifact*. In addition, we extended the definition of adaptability. As revealed by the coding of the 121 papers, this term sometimes refers to dynamic reaction to changing environmental conditions. Since we had omitted this meaning, we completed the definition of the criterion as follows: *the ease with which the artifact can work in contexts other than those for which it was specifically designed, or change according to evolutions in context*.

Improving Simplicity

We simplified the hierarchy by removing some of the seventeen criteria not found in our sample. Various reasons might explain why a criterion, suggested by the IS literature on artifact evaluation, was not found in the sample of 121 papers. This required care in deciding which criteria should be kept or removed.

As mentioned above, ethicality and side effects are crucial in assessing the impact of artifacts. We should keep them in the hierarchy. The DSR community is urged to define evaluation methods for assessing them. The criteria pertaining to consistency, alignment, and fit are related. They often appear in the design science literature and should be kept in the hierarchy. That alignment with business was not found among the criteria in our sample is a sign of partial disconnect between DSR and organization-wide concerns. Finally, the criteria of modifiability and structure/simplicity, although not found in the sample, apply to many types of artifacts, including taxonomies (as this paper illustrates). We should keep them in the hierarchy.

We remove the other criteria from the hierarchy. Despite early suggestions to consider the style (aka elegance) of artifacts [26, 34], design science researchers

have shown little interest in this criterion. Today, evaluating the impact of artifacts appears much more critical than assessing their style. Concerning construct overload, redundancy, excess, and deficit, they should actually be considered as measures of correspondence with another model, instead of criteria. Inside the category "activity," the fact that functionality was never assessed suggests that it is not a specific category. It is a specific case of activity/completeness. Finally, the coding of the 121 papers revealed the difficulty of distinguishing efficiency from performance. We remove efficiency from the hierarchy of criteria, and redefine performance as follows: *the degree to which the artifact accomplishes its function with given constraints of resources*. Time and space are specific cases of resources.

Figure 2 shows the revised hierarchy of criteria. To conclude the taxonomy development process, we evaluate the taxonomy summatively. This includes the

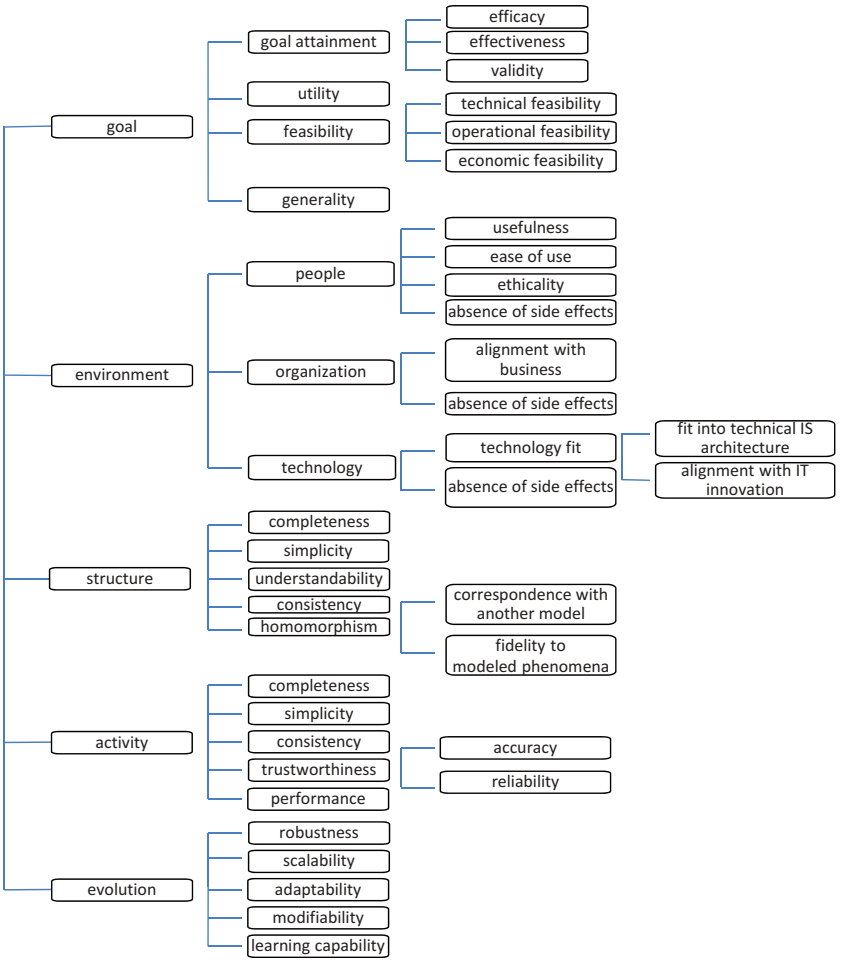

Figure 2. Final Hierarchy of Evaluation Criteria

evaluation of simplicity to assess the impact of the revisions to the hierarchy of criteria.

Summative Evaluation of the Taxonomy

Three subjective ending conditions are relevant here: the taxonomy should be concise, robust, and explanatory. "Concise" is the equivalent of structure/simplicity in our hierarchy of criteria. Applying the metric of online Supplement B to the final taxonomy yields a simplicity value of 0.22 (compared with 0.21 after the conceptualization phase). The simplicity of the hierarchy of evaluation criteria is 0.27 (compared with 0.26). Thus, the revisions to the taxonomy and, more specifically, the revisions to the hierarchy of criteria, have only slightly improved simplicity. One reason is that additions of new characteristics partly offset simplifications. This also suggests that our simplicity metric is conservative: even though the net effect of the revisions to the hierarchy was the removal of five characteristics, the impact on the computed simplicity was limited.

"Robust" means that the taxonomy should provide differentiation among objects, that is, the expression of subtle differences between objects inasmuch as these differences are relevant. This criterion is specific to taxonomies; it has no equivalent in our hierarchy of criteria. In this work, we needed to represent different aspects of artificiality and naturalness in evaluation methods. The taxonomy enables the expression of nuances in the distinction between artificial and naturalistic evaluation. For example, considering the dimensions "evaluation technique," "level of evaluation," and "secondary participants," the following combinations of characteristics represent gradual increase in naturalness: (*illustrative scenario, fictitious example or examples, students*), (*illustrative scenario, real example or examples, students*), (*illustrative scenario, real example or examples, practitioners*), and (*case study, real example or examples, practitioners*). The taxonomy also enables the expression of nuances in the "what" of evaluation, based on the hierarchy of criteria. For example, the hierarchy distinguishes five criteria pertaining to evolution.

"Explanatory" means that the taxonomy should provide useful explanations on the nature of classified objects. To assess the utility of the taxonomy, we compare this research with Venable et al. [58] (to the best of our knowledge, this is the only paper from the AIS basket dedicated to evaluation in DSR). These two works appear to be complementary on several aspects. We address the tactical and operational levels of IS artifact evaluation, versus the strategic level in Venable et al. [58]. We focus on the "what" and the "how" of evaluation, versus a high-level view of the "why," the "when," the "what," and the "how." We define a hierarchy of universal criteria for evaluation in DSR, versus considering the definition of universal criteria as problematic. We exhibit prototypical evaluation methods (compositional styles) for IS artifacts, versus prototypical evaluation strategies.

Conclusion

The contributions of this paper are twofold. The first contribution is the taxonomy of IS artifact evaluation methods. This taxonomy specifies the dimensions pertaining to the "what" and the "how" of evaluation, thereby answering questions *RQ1* and *RQ2*. Regarding the "what*"* of evaluation, we view the output of a DSR endeavor as a system of IS artifacts, and structure the hierarchy of evaluation criteria accordingly. Our systems approach provides the holistic view that has been missing so far in IS artifact evaluation. Regarding the "how," we specify the different dimensions of evaluation methods. We built and evaluated our taxonomy according to the DSR paradigm. We developed it by conceptual analysis of the design science literature, followed by empirical content analysis of a sample of design research papers. To evaluate the taxonomy, we applied some criteria from our hierarchy of criteria.

The results of the empirical analysis of 121 DSR papers represent our second contribution, addressing question *RQ3*. More specifically, this analysis improves understanding of the relationships between different dimensions of evaluation, identifies typical patterns ("compositional styles") in IS artifact evaluation, and brings to light unexplored criteria. We distill the results of the empirical analysis into four evaluation guidelines for design science researchers.

Guideline 1: Consider commonly used compositional styles in IS artifact evaluation. The seven most common styles are (1) demonstration, (2) simulation- and metric-based benchmarking of artifacts, (3) practice-based evaluation of effectiveness, (4) simulation- and metric-based absolute evaluation of artifacts, (5) practice-based evaluation of usefulness or ease of use, (6) laboratory, student-based evaluation of usefulness, and (7) algorithmic complexity analysis. These styles may be viewed as reusable evaluation methods. The choice of the appropriate style depends on the criterion to evaluate and on the other dimensions of evaluation.

Guideline 2: Generate new evaluation methods creatively and pragmatically, considering relationships between and within the "what" and the "how" of evaluation. Beyond reusing common evaluation methods, design science researchers are also encouraged to generate new methods. The variety of methods is already noticeable in practice (e.g., regarding the evaluation of usefulness). When generating a method, the relationships between the "what" and the "how" of evaluation, and between the different dimensions of the "how," should be considered. The assessed criterion influences the choice of method, more specifically the evaluation technique and the form and relativeness of evaluation. Moreover, the association rules in Table 7 illustrate some interdependencies between evaluands (IS artifacts composing the evaluated system), evaluation criteria, and different dimensions in the "how" of evaluation (e.g., rule R4 relates three dimensions in the "how" of evaluation). Even if the "how" of evaluation is generally chosen depending on the "what" (e.g., evaluation methods are chosen according to the criteria to evaluate), pragmatic considerations (e.g., the unavailability of secondary participants) may also influence the choice of the criteria to evaluate. Thus, there are several possible ways of interpreting the interdependencies between and within the "what" and the "how"

of evaluation. In generating new evaluation methods, creativity is encouraged. For example, to evaluate criteria pertaining to the structure of artifacts, software engineering metrics may be adapted.

Guideline 3: Investigate the organizational impact of IS artifacts. Judging from our sample of papers, organizational impact, the ultimate measure of IS success [15], is neglected in artifact evaluation: in the hierarchy of criteria, no criterion pertaining to environment/organization is assessed. The focus is on individual impact (usefulness). The link with organizational impact is missing. As Gill and Hevner [17] claim, and as our data confirm, DSR has thus far overemphasized immediate utility, at the expense of sustainable impact. Assessment of the long-term impact of artifacts on organizations (including economic impact) should be encouraged.

Guideline 4: Evaluate the societal impact (including ethical considerations) of IS artifacts, when relevant. Societal impact should be a major concern of artifact evaluation [9, 40]. However, according to our data, it is not measured in practice. IS researchers may innovate by evaluating the potential impact of their artifacts on society. This requires the development of new evaluation methods, for example, to measure ethicality and side effects.

Our findings, and the characteristics and dimensions of the taxonomy of evaluation methods, should be interpreted in light of the limitations of this work. First, this research focuses on the evaluation of the output of DSR. It does not consider the evaluation of the DSR process itself. Second, we could customize the hierarchy of criteria to reflect the specificities of certain artifacts, for example, design theories. Third, beyond the "how" and the "what," our taxonomy could also consider other aspects of artifact evaluation that are often implicit in published research, for example, the "why" [57]. Fourth, in the hierarchy of criteria, the three categories of environment [26] could be enriched to include other categories of stakeholders, for example, society at large. Fifth, despite the reasonable sample size, this study should be replicated to papers from other journals, for example, *Decision Support Systems*. Finally, based on a larger sample, we could investigate differences between major IS journals as regards artifact evaluation practice. This will be the object of future work. We also plan to develop metrics, and more generally evaluation methods, for assessing the ethicality and side effects of artifacts.

NOTE

1. The term "compositional style", commonly used in music, was suggested by one reviewer. It has already been applied in IS [35]. In art, composition is "the harmonious arrangement of the parts of a work [...] in relation to each other and to the whole" (collinsdictionary.com). Compositional styles in IS artifact evaluation are typical ways of assembling the six dimensions of the taxonomy to form evaluation methods.

Supplemental Data

Supplemental data for this article can be accessed on the publisher's website at http://dx.doi.org/10.1080/07421222.2015.1099390

References

1. Abbasi, A.; Zahedi, F.M.; Zeng, D.; Chen, Y.; Chen, H.; and Nunamaker, J.F. Enhancing predictive analytics for anti-phishing by exploiting website genre information. *Journal of Management Information Systems, 31*, 4 (Spring 2015), 109–157.
2. Ackoff, R.L. Towards a system of systems concepts. *Management Science*, *17*, 11 (1971), 661–671.
3. Aier, S., and Fischer, C. Criteria of progress for information systems design theories. *Information Systems and e-Business Management*, *9*, 1 (2011), 133–172.
4. Arnott, D., and Pervan, G. Design science in decision support systems research: An assessment using the Hevner, March, Park, and Ram guidelines. *Journal of the Association for Information Systems*, *13*, 11 (2012), 923–949.
5. Bergman, M.; Lyytinen, K.; and Mark, G. Boundary objects in design: An ecological view of design artifacts. *Journal of the Association for Information Systems*, *8*, 11 (2007), 546–568.
6. Bondi, A.B. Characteristics of scalability and their impact on performance. *Proceedings of the Second International Workshop on Software and Performance* (WOSP 2000). Ottawa: ACM, 2000, pp. 195–203.
7. Bowler, T. *General Systems Thinking: Its Scope and Applicability.* New York: North Holland, 1981.
8. Checkland, P., and Scholes, J. *Soft Systems Methodology in Action*. Chichester: Wiley, 1990.
9. Chen, H. Editorial: Design science, grand challenges, and societal impacts. *ACM Transactions on Management Information Systems*, *2*, 1 (2011), 1:1–1:10.
10. Churchman, C.W. *The Design of Inquiring Systems: Basic Concepts of Systems and Organization*. New York: Basic Books, 1971.
11. Cleven, A.; Gubler, P.; and Hüner, K.M. Design alternatives for the evaluation of design science research artifacts. In V. Vaishnavi and S. Purao (eds.), *Proceedings of the Fourth International Conference on Design Science Research in Information Systems and Technology* (DESRIST 2009). Philadelphia, PA: ACM, 2009, pp. 1–8.
12. Coners, A., and Matthies, B. A content analysis of content analyses in IS research: Purposes, data sources, and methodological characteristics. In K. Siau, Q. Li, and X. Guo (eds.), *Proceedings of the Eighteenth Pacific Asia Conference on Information Systems* (PACIS 2014)/ Chengdu, China: Association for Information Systems, 2014, pp. 1–16.
13. Creighton, C., and Hanash, S. Mining gene expression databases for association rules. *Bioinformatics*, *19*, 1 (2003), 79–86.
14. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*, 3 (1989), 319–340.
15. DeLone, W.H., and McLean, E.R. Information systems success: The quest for the dependent variable. *Information Systems Research*, *3*, 1 (1992), 60–95.
16. Fischer, C. The information systems design science research body of knowledge: A citation analysis in recent top-journal publications. In P.B. Seddon and S. Gregor (eds.), *Proceedings of the Fifteenth Pacific Asia Conference on Information Systems* (PACIS 2011)/ Brisbane: Association for Information Systems, 2011, pp. 1–12.
17. Gill, T.G., and Hevner, A.R. A fitness-utility model for design science research. *ACM Transactions on Management Information Systems*, *4*, 2 (2013), 1–24.
18. Glinz, M. On non-functional requirements. In A. Sutcliffe and P. Jalote (eds.), *Proceedings of the Fifteenth IEEE International Requirements Engineering Conference* (RE 2007). New Delhi: IEEE, 2007, pp. 21–26.
19. Gregor, S. The nature of theory in information systems. *MIS Quarterly*, *30*, 3 (2006), 611–642.
20. Gregor, S., and Hevner, A.R. Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*, 2 (2013), 337–355.
21. Gregor, S., and Iivari, J. Designing for mutability in information systems artifacts. In D. Hart and S. Gregor (eds.), *Information Systems Foundations: Theory, Representation and Reality.* Canberra: Australian National University Press, 2007, pp. 3–24.
22. Gregor, S., and Jones, D. The anatomy of a design theory. *Journal of the Association for Information Systems*, *8*, 5 (2007), 312–335.

23. Hahsler, M.; Buchta, C.; Grün, B.; and Hornik, K. *arules: Mining association rules and frequent itemsets*, R package version 1.1-6, http://CRAN.R-project.org/package=arules, 2014.
24. Hair, J.; Black, W.; Babin, B.; and Anderson, R. *Multivariate Data Analysis: A Global Perspective*. 7th ed. Upper Saddle River, NJ: Hall, 2010.
25. Henderson, J.C., and Venkatraman, N. Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal*, *32*, 1 (1993), 4–16.
26. Hevner, A.R.; March, S.T.; Park, J.; and Ram, S. Design science in information systems research. *MIS Quarterly*, *28*, 1 (2004), 75–105.
27. Huang, C.-Y., and Kuo, S.-Y. Analysis of incorporating logistic testing-effort function into software reliability modeling. *IEEE Transactions on Reliability*, *51*, 3 (2002), 261–270.
28. Huysmans, P., and De Bruyn, P. A mixed methods approach to combining behavioral and design research methods in information systems research. *Proceedings of the Twenty-First European Conference on Information Systems* (ECIS 2013). Utrecht, the Netherlands: Association for Information Systems, 2013, pp. 1–12.
29. ISO/IEC/IEEE, *Systems and Software Engineering: Vocabulary*, ISO/IEC/IEEE 24765:2010(E), December 2010, pp. 1–418.
30. Le Moigne, J.-L. Modeling for reasoning socio-economic behaviors. *Cybernetics and Human Knowing*, *13*, 3–4 (2006), 9–26.
31. Liu, F., and Myers, M.D. An analysis of the AIS basket of top journals. *Journal of Systems and Information Technology*, *13*, 1 (2011), 5–24.
32. Lu, M.-T., and Yeung, W.-L. A framework for effective commercial web application development. *Internet Research*, *8*, 2 (1998), 166–173.
33. Lukyanenko, R., and Parsons, J. Reconciling theories with design choices in design science research. In J. vom Brocke, R. Hekkala, S. Ram, and M. Rossi (eds.), *Proceedings of the Eighth International Conference on Design Science Research in Information Systems and Technology* (DESRIST 2013). Helsinki: Springer Verlag, 2013, pp. 165–180.
34. March, S.T., and Smith, G.F. Design and natural science research on information technology. *Decision Support Systems*, *15*, 4 (1995), 251–266.
35. Mathiassen, L.; Chiasson, M.; and Germonprez, M., Compositional styles in action research: A critical analysis of leading information systems journals. *Sprouts: Working Papers on Information Systems*, *9*, 35 (2009), Lancaster University, UK. http://sprouts.aisnet.org/9-35/.
36. Moore, G., and Benbasat, I. Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research*, *2*, 3 (1991), 192–222.
37. Myers, M.D., and Venable, J.R. A set of ethical principles for design science research in information systems. *Information and Management*, *51*, 6 (2014), 801–809.
38. Nickerson, R.C.; Varshney, U.; and Muntermann, J. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22*, 3 (2013), 336–359.
39. Nunamaker, J.F.; Chen, M.; and Purdin, T.D.M. Systems development in information systems research. *Journal of Management Information Systems*, *7*, 3 (1990–91), 89–106.
40. Nunamaker, J.F.; Derrick, D.C.; Elkins, A.C.; Burgoon, J.K.; and Patton, M.W. Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems*, *28*, 1 (2011), 17–48.
41. Offermann, P.; Blom, S.; Schönherr, M.; and Bub, U. Artifact types in information systems design science: A literature review. In R. Winter, J.L. Zhao, and S. Aier (eds.), *Proceedings of the Fifth International Conference on Design Science Research in Information Systems and Technology* (DESRIST 2010). St. Gallen: Springer Verlag, 2010, pp. 77–92.
42. Orlikowski, W.J., and Iacono, C.S. Research commentary: Desperately seeking the "IT" in IT research: A call to theorizing the IT artifact. *Information Systems Research*, *12*, 2 (2001), 121–134.
43. Peffers, K.; Rothenberger, M.; Tuunanen, T.; and Vaezi, R. Design science research evaluation. In K. Peffers, M. Rothenberger, and B. Kuechler (eds.), *Proceedings of the Seventh International Conference on Design Science Research in Information Systems and Technology* (DESRIST 2012). Las Vegas: Springer Verlag, 2012, pp. 398–410.

44. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; and Chatterjee, S. A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*, 3 (2007–8), 45–77.

45. R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2015. http://www.R-project.org.

46. Sangupamba Mwilu, O.; Prat, N.; and Comyn-Wattiau, I. Business intelligence and big data in the cloud: oppportunities for design-science researchers. In M. Indulska and S. Purao (eds.), *Proceedings of ER 2014 Workshops*. Atlanta, GA: Springer International, 2014, pp. 75–84.

47. Siau, K., and Rossi, M. Evaluation techniques for systems analysis and design modelling methods: A review and comparative analysis. *Information Systems Journal*, *21*, 3 (2011), 249–268.

48. Simon, H. *The Sciences of the Artificial*. 3rd ed. Cambridge, MA: MIT Press, 1996.

49. Skyttner, L. *General Systems Theory: Problems, Perspectives, Practice*. 2nd ed. Singapore: World Scientific, 2005.

50. Sonnenberg, C., and vom Brocke, J. Evaluations in the science of the artificial: Reconsidering the build-evaluate pattern in design science research. In K. Peffers, M. Rothenberger, and B. Kuechler (eds.), *Proceedings of the Seventh International Conference on Design Science Research in Information Systems and Technology* (DESRIST 2012). Las Vegas: Springer Verlag, 2012, pp. 381–397.

51. Spanoudakis, G., and Constantopoulos, P. Elaborating analogies from conceptual models. *International Journal of Intelligent Systems*, *11*, 11 (1996), 917–974.

52. Sun, Y., and Kantor, P.B. Cross-evaluation: A new model for information system evaluation. *Journal of the American Society for Information Science and Technology*, *57*, 5 (2006), 614–628.

53. Tremblay, M.C.; Hevner, A.R.; and Berndt, D.J. Focus groups for artifact refinement and evaluation in design research. *Communications of the Association for Information Systems*, *26*, 1 (2010), 599–618.

54. Twyman, N.W.; Lowry, P.B.; Burgoon, J.K.; and Nunamaker, J.F. Autonomous scientifically controlled screening systems for detecting information purposely concealed by individuals. *Journal of Management Information Systems*, *31*, 3 (2014), 106–137.

55. Vaishnavi, V., and Kuechler, B., *Design Science Research in Information Systems*. http://desrist.org/design-research-in-information-systems, January 2004 (last updated October 2013), pp. 1–45.

56. Venable, J. Rethinking design theory in information systems. In J. vom Brocke, R. Hekkala, S. Ram, and M. Rossi (eds.), *Proceedings of the Eighth International Conference on Design Science Research in Information Systems and Technology* (DESRIST 2013). Helsinki: Springer Verlag, 2013, pp. 136–149.

57. Venable, J.; Pries-Heje, J.; and Baskerville, R. A comprehensive framework for evaluation in design science research. In K. Peffers, M. Rothenberger, and B. Kuechler (eds.), *Proceedings of the Seventh International Conference on Design Science Research in Information Systems and Technology* (DESRIST 2012). Las Vegas: Springer Verlag, 2012, pp. 423–438.

58. Venable, J.; Pries-Heje, J.; and Baskerville, R. FEDS: A framework for evaluation in design science research. *European Journal of Information Systems* (advance online publication, November 2014), doi:10.1057/ejis.2014.36.

59. von Bertalanffy, L. *General System Theory: Foundations, Development, Applications*. Rev. ed. New York: George Braziller, 1969.

60. Walls, J.G.; Widmeyer, G.R.; and El Sawy, O.A. Building an information system design theory for vigilant EIS. *Information Systems Research*, *3*, 1 (1992), 36–59.

61. Wand, Y., and Weber, R. On the ontological expressiveness of information systems analysis and design grammars. *Information Systems Journal*, *3*, 4 (1993), 217–237.

62. Wang, R., and Strong, D. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, *12*, 4 (1996), 5–33.

63. Wang, S., and Wang, H. Towards innovative design research in information systems. *Journal of Computer Information Systems*, *51*, 1 (2010), 11–18.

64. Weber, R.P. *Basic Content Analysis*. 2nd ed. Newbury Park, CA: Sage, 1990.
65. Weinberg, G. *An Introduction to General Systems Thinking*. New York: Wiley, 1975.
66. Wieringa, R. Tutorial #2: Empirical validation research methods. *Communication at the Sixth International Conference on Research Challenges in Information Science* (RCIS 2012). Valencia, Spain, May 2012.
67. Williams, K.; Chatterjee, S.; and Rossi, M. Design of emerging digital services: a taxonomy. *European Journal of Information Systems*, *17*, 5 (2008), 505–517.
68. Winter, R. Design science research in Europe. *European Journal of Information Systems*, *17*, 5 (2008), 470–475.
69. Zlotkin, G., and Rosenschein, J.S. Cooperation and conflict resolution via negotiation among autonomous agents in non-cooperative domains. *IEEE Transactions on Systems, Man and Cybernetics*, *21*, 6 (1991), 1317–1324.

Appendix: Coding Scheme

A. Global coding for the paper
   1. The design artifact
      1.1. Type of artifact **(multiple choices allowed)**...
      1.2. Brief description of the artifacts
   2. Additional criteria (if any) (assessed in the paper, but absent from the hierarchy of criteria in 6) **(optional)**

B. Coding for each evaluation method
   3. Brief description of the evaluation method (**for a measure, specify the name of the metric**)
   4. Evaluation technique ...
   5. Form of evaluation ...
   6. Assessed criterion

Goal

Goal attainment

1. Efficacy: *The degree to which the artifact achieves its goal considered narrowly, without addressing situational concerns [8, 57]*.
2. Effectiveness: *The degree to which the artifact achieves its goal in a real situation [8, 57]*.
3. Validity: *Validity means that the artifact works correctly, i.e. correctly achieves its goal [20]*.
4. Utility: *Utility measures the value of achieving the artifact's goal, i.e. the difference between the worth of achieving this goal and the price paid for achieving it [20, 69]*.

Feasibility

5. Technical feasibility: *Evaluates, from a technical point of view, the ease with which a proposed artifact will be built and operated [5, 32]*.
6. Operational feasibility: *Evaluates the degree to which management, employees, and other stakeholders, will support the proposed artifact, operate it, and integrate it into their daily practice [5, 32]*.

(continues)

Continued

7. Economic feasibility: *Evaluates whether the benefits of a proposed artifact would outweigh the costs of building and operating the artifact [5, 32]*.
8. Generality: *Refers to the scope of the artifact's goal. The broader the goal scope, the more general the artifact [3, 22]*.

Environment

People

9. Usefulness: *The degree to which the artifact positively impacts the task performance of individuals [14]*.
10. Ease of use: *The degree to which the use of the artifact by individuals is free of effort [14]*.
11. Ethicality: *The degree to which the artifact complies with ethical principles*.
12. Absence of side effects: *The degree to which the artifact is free of undesirable impacts on individuals in the long run [57]*.

Organization

13. Alignment with business: *The congruence of the artifact with the organization and its strategy [25]*.
14. Absence of side effects: *The degree to which the artifact is free of undesirable impacts on the organization in the long run [57]*.

Technology

Technology fit

15. Fit into technical IS architecture: *The degree to which the artifact integrates into the technical IS architecture of the organization*.
16. Alignment with IT Innovation: *The degree to which the artifact uses innovative IT [63]*.
17. Absence of side effects: *The degree to which the artifact is free of undesirable impacts on the technical IS architecture of the organization in the long run [57]*.

Structure

18. Completeness: *The degree to which the structure of the artifact contains all necessary elements and relationships between elements*.
19. Simplicity: *The degree to which the structure of the artifact contains the minimal number of elements and relationships between elements [29]*.
20. Style: *The elegance with which the artifact has been built* [26, 34].
21. Consistency: *The degree of uniformity, standardization, and freedom from contradiction among the elements of the structure of the artifact [29]*.

Homomorphism

Correspondence with another model: *The degree to which the structure of the artifact corresponds to a reference model*.

22. Construct overload: *Construct overload occurs when one construct in the structure of the artifact maps to two or more constructs in the reference model [61]*.
23. Construct redundancy: *Construct redundancy occurs when two or more constructs in the structure of the artifact are used to represent a single construct in the reference model [61]*.

(continues)

Continued

24. Construct excess: *Construct excess occurs when one construct in the structure of the artifact does not map to any construct in the reference model [61]*.
25. Construct deficit: *Construct deficit occurs when one construct in the reference model does not map to any construct in the structure of the artifact [61]*.
26. Fidelity to modeled phenomena: *The degree to which the structure of the artifact corresponds to the modeled reality*.

Activity

27. Completeness: *The degree to which the activity of the artifact contains all necessary elements and relationships between elements*.
28. Functionality: *The capability of the artifact to provide functions which meet stated and implied needs when the artifact is used under specified conditions [29]*.
29. Simplicity: *The degree to which the activity of the artifact contains the minimal number of elements and relationships between elements [29]*.
30. Consistency: *The degree of uniformity, standardization, and freedom from contradiction among the elements of the activity of the artifact [29]*.

Trustworthiness

31. Accuracy: *The degree of agreement between outputs of the artifact and the expected outputs [29]*.
32. Reliability: *The ability of the artifact to function correctly in a given environment during a specified period of time [27]*.
33. Performance: *The degree to which the artifact accomplishes its functions within given constraints of time or space. Speed and throughput (the amount of output produced in a given period of time) are examples of time constraints. Memory usage is an example of space constraint [18, 29]*.
34. Efficiency: *The maximization of the ratio between outputs and inputs of the artifact*.

Evolution

35. Robustness: *The ability of the artifact to handle invalid inputs or stressful environmental conditions [29]*.
36. Scalability: *The ability of the artifact to either handle growing amounts of work in a graceful manner, or to be readily enlarged [6, 67]*.
37. Adaptability: *The ease with which the artifact can work in contexts other than those for which it was specifically designed. Synonym: flexibility [29]*.
38. Modifiability: *The ease with which the artifact can be changed without introducing defects [29]*.
39. Learning capability: *The ability of the artifact to learn from experience*.

7. Secondary participants ...
8. Level of evaluation ...
9. Relativeness of evaluation ...