Bus Inf Syst Eng 58(6):397–414 (2016)
DOI 10.1007/s12599-016-0427-3

RESEARCH PAPER

Evaluation Methods in Process-Aware Information Systems Research with a Perspective on Human Orientation

Simone Kriglstein · Maria Leitner · Sonja Kabicher-Fuchs · Stefanie Rinderle-Ma

Published online: 1 March 2016

Abstract Research on process-aware information systems (PAIS) has experienced a dramatic growth in recent years. Lately, a particular increase of empirical studies and focus on human oriented research questions could be observed, leading to an expansion of applied evaluation methods in PAIS research. At the same time, it can be observed that evaluation methods are not always applied in a systematic manner and related terminology is at times used in an ambiguous way. Hence, the paper aims at investigating evaluation methods that are typically employed in PAIS research with a special focus on human orientation. The applied methodology includes a literature review, an expert survey, and a focus group. The authors present their findings as a collection of typical evaluation methods and the related PAIS artifacts. They highlight which evaluation methods are currently used and which evaluation methods could be of interest for future PAIS research efforts.

Keywords Evaluation · Human orientation · Process-aware information systems · Security · Visualization

Accepted after two revisions by Prof. Dr. Becker.

Electronic supplementary material The online version of this article (doi:10.1007/s12599-016-0427-3) contains supplementary material, which is available to authorized users.

Dr. S. Kriglstein (✉)
Institute for Design and Assessment of Technology, Faculty of Informatics, Vienna University of Technology, Argentinierstraße 8, 2. Stock, 1040 Vienna, Austria
e-mail: simone.kriglstein@tuwien.ac.at

Dr. M. Leitner
Digital Safety and Security Department, Information Management, AIT Austrian Institute of Technology GmbH, Donau-City-Straße 1, 1220 Vienna, Austria
e-mail: maria.leitner@ait.ac.at

Dr. S. Kabicher-Fuchs · Univ.-Prof. Dr. S. Rinderle-Ma
Faculty of Computer Science, University of Vienna, Währingerstraße 29, 1090 Vienna, Austria
e-mail: sonja.kabicher-fuchs@univie.ac.at

Univ.-Prof. Dr. S. Rinderle-Ma
e-mail: stefanie.rinderle-ma@univie.ac.at

1 Introduction

Process-aware information systems (PAIS) have been an intensively discussed topic in practice and in science since the late 1970's (e.g., Dumas et al. 2005; van der Aalst 2013). Since then the importance of PAIS has increased steadily and resulted in a number of PAIS-related tools (e.g., workflow systems, business process management suites, and business process modeling tools), in the practical application of the concepts and the systems, and in a continuously growing body of professional and scientific literature (cf. Dumas et al. 2005).

1.1 Problem Statement

We perceive PAIS as a type of information system "that manages and executes operational processes involving people, applications, and/or information sources on the basis of process models" (Dumas et al. 2005). PAIS in practice have been considered as drivers for companies' benefits such as (cf. Vanderfeesten and Reijers 2005): the specification, execution and control of business processes, easy coordination of work, higher quality of services, efficient execution of work, and flexible processes. The support of their processes through PAIS pervades nearly

any application area or industry such as manufacturing (Schulte et al. 2012),[1] health care (Lenz and Reichert 2007), and finance (cf. Rabobank processes of BPI'14 challenge, http://www.win.tue.nl/bpi/2014/challenge). Although PAIS are meant to assist human workers (cf. Klein et al. 2000), such systems have received criticism as well (cf. Vanderfeesten and Reijers 2005): particularly work psychologists and potential (future) users fear that workflow systems could lead to mechanical, monotonous, and rigid office work (Kabicher-Fuchs et al. 2013) and to the perception that system users are exchangeable resources. Several papers have addressed different aspects of PAIS that seem to be particularly relevant for the human orientation of such systems and for successful human computer interaction.

In the following, it is claimed that human orientation in PAIS constitutes an important challenge and is of increasing interest for the research community. This claim is supported by the broader literature on software systems and information systems (IS) (Gediga and Hamborg 2001; Irani 2002), by PAIS specific literature (Kabicher-Fuchs et al. 2012, 2013; Kabicher-Fuchs and Rinderle-Ma 2012), and by the recent advent of related topics in IS conferences.[2] Starting from this, a further claim is that research on human orientation requires special evaluation methods and artifacts. As stated in Song and Letch (2012) "[t]he study of human factors in evaluation also consists with the shift from traditional evaluation to understanding-driven stream". Another study (Serafeimidis and Smithson 2003) claims that "[t]he organizational and subjective nature of IS evaluation brings into the foreground the human actor, a stakeholder, of an evaluation exercise". Following both claims, the motivation for the study at hand is to investigate evaluation methods and artifacts that have been used in PAIS with specific focus on human orientation. We understand human orientation in PAIS in a rather general sense, i.e., as an umbrella term for humans playing a role in a PAIS (cf. Vanderfeesten and Reijers 2005; Kabicher-Fuchs et al. 2012).

In scientific literature, several analyses of research in the context of PAIS, and more generally of IS, have been conducted (see, e.g., Glass et al. 2002; Hevner et al. 2004; Ramesh et al. 2004; Wilde and Hess 2007; Houy et al. 2010). Such analyses create a firm foundation for advancing knowledge and for an informed understanding of existing research and gaps where further work and research is needed. However, according to the analysis in Song and Letch (2012), in the most recent time period considered in the study (i.e., 2006–2010), only one construct/measurement validation was conducted in IS. Specifically, a systematic analysis of evaluation methods as applied in PAIS in the context of human orientation is entirely missing.

In this work, we investigate which evaluation methods and artifacts have been used in PAIS research with focus on human orientation in general and graphical presentations (visualization) but also on security topics, since security can fail due to misunderstandings, wrong communication or false assumptions between users and systems. In order to illustrate the terms 'artifact' and 'evaluation method' in this context, take the paper *Visual change tracking for business process models* (Kabicher et al. 2011) as an example. Here, artifacts to be evaluated are graphical descriptions of process models, and questionnaires are used as evaluation method for this empirical study. Research in the IS field is characterized by the research paradigms 'behavioral science' and 'design science' (cf. Houy et al. 2010). Whereas Houy et al. (2010) analyzed empirical research following the behavioral and design science research paradigm in business process management (BPM), we have in our review concentrated on studies which investigated human oriented factors in the field of PAIS by following a design science research approach.

This study reviews scientific contributions that focus on human orientation in general (e.g., user, function, and task allocation), visualization (e.g., graphical representation of process models, and task lists), and security (e.g., access control, and privileges) in PAIS. We categorize the identified PAIS artifacts into theoretical and executable artifacts and classify the evaluation methods according to the methodical framework presented by Gediga and Hamborg (2001) into the categories *Behavior-based*, *Opinion-based*, and *Predictive* evaluation methods. The paper centers on the three following research questions: Which evaluation methods have been used so far to examine PAIS artifacts that are in direct contact with users? Which evaluation methods are of future interest? Can these evaluation methods be classified into the categories mentioned above?

1.2 Procedure

We proceed based on the following methodology. In a first step, a literature review is conducted in order to identify which artifacts and evaluation methods have so far been

[1] See especially the demand for adequate process support in Industrie 4.0 or Factories of the Future programs of the H2020 program (http://ec.europa.eu/research/industrial_technologies/factories-of-the-future_en.html).

[2] Major conferences in the IS and PAIS area list related topics, e.g., *"User-centric aspects of BPM"* and *"Human-centric processes and knowledge-intensive processes"* in BPM 2015 CfP (http://bpm2015.q-e.at/call-for-contributions/call-for-papers/), *"... systems [...] appealing to large and diverse user bases"* in CaISE 2015 CfP (http://caise2015.dsv.su.se/call-for-papers/), *"Human-centred Information Systems"* in ECIS 2015 CfP (http://www.ecis2015.eu/participation/call-for-papers).

applied in scientific contributions addressing human orientation of PAIS. Moreover, the relationship between the identified artifacts and evaluation methods is analyzed. The literature review yields an overview of the 'as-is' state concerning evaluation methods used with respect to human orientation in PAIS.

After the literature review, an expert survey shall provide insights into experts' awareness concerning artifacts, evaluation methods, and their relationships, that have been used so far in research on human oriented aspects of PAIS. Further on, the expert survey shall also highlight the experts' forecasts concerning artifacts, evaluation methods, and their relationships, that will be of future interest in the field of human orientation of PAIS.

Finally, a focus group supplements the study with further experts' opinions concerning artifacts, evaluation methods, and their relationships used so far and with future potential to support and investigate human orientation of PAIS.

Although we gained valuable insights from the expert survey, we also identified misunderstandings about the meanings of questions or contradictory responses. Therefore, we decided to conduct a focus group session (cf. Stewart et al. 2007) which is a good method to analyze the results from different points of view in a short period of time. In particular, the focus group session gave us the possibility to discuss and verify the results, from the literature review (RQ1) and, from the expert survey.

Both, the expert survey and the focus group are intended to shed light on the 'to-be' state of evaluation methods used with respect to human orientation in PAIS. We decided to conduct a focus group session additionally to the expert survey to discuss and verify the results of the expert survey, since we identified misunderstandings concerning the meanings of questions or contradictory responses within the expert survey. Of particular interest is the comparison of 'as-is' and 'to-be' state as it might reveal blank spots in the evaluation landscape. The objectives of the study are threefold. First of all, it identifies and categorizes artifacts that have been developed in the context of human orientation of PAIS. In addition, evaluation methods are identified which have been used to evaluate artifacts that have been developed in the context of human orientation of PAIS. These evaluation methods are also evaluated according to the theoretical methodical framework as proposed by Gediga and Hamborg (2001).

1.3 Contribution

This work presents an analysis of artifacts and evaluation methods which have been used in PAIS research with a focus on aspects that are particularly relevant for humans working with the system such as PAIS users and employees. Based on the findings of the literature review, expert survey and the focus group, we highlight ten widely used evaluation methods and interdisciplinary evaluation methods (e.g., combination human orientation and security, human orientation and visualization, as well as security and visualization). Furthermore, we discuss the trend towards human orientation and give recommendations for using evaluation methods in PAIS as well as an overview of evaluation methods and their use in human orientation, security, and visualization.

This article is structured as follows: Sect. 2 introduces the basic definition and concepts followed by an overview on related work and the classification applied in the remainder of the paper. Section 3 describes the overarching methodology used in the paper. This is followed by a literature review in Sect. 4, an expert survey in Sect. 5, and a focus group in Sect. 6. The results of the literature review, of the expert survey, and of the focus group are summarized in Sect. 7. The discussion provided in Sect. 8 includes ten widely used evaluation methods, interdisciplinary evaluation methods, recommendations, and comments on lessons learned. It furthermore discusses limitations of the paper. Section 9 concludes the paper.

2 Background: Evaluation Methods in PAIS

In this section, we will explain the concepts and definitions used in the rest of the paper. Subsequently, we will discuss how, in general, evaluation methods are used in IS and related areas in order to put the study into a broader context. Finally, the classification of evaluation methods that is utilized in the study is introduced.

2.1 Concepts and Definitions

Figure 1 illustrates the main concepts *Artifact*, *Type of Investigation*, *Strategy*, and *Evaluation Method* as used in the PAIS research papers analyzed in this article (presented in detail in the literature review in Sect. 4). The type of investigation determines the strategy and artifact selected for in the evaluation. The strategy and artifact, in turn, determine the evaluation method.

Which kind of investigation is chosen (e.g., case study or user evaluation) depends strongly on the purpose and goal of the investigation itself, for instance to determine usability or security aspects (e.g., if the goal is to observe how users will work with the system, user evaluation is a promising way). According to the type of investigation, different evaluation methods can be used (e.g., thinking aloud) and can be applied for all items (i.e., artifacts) that are produced during the development process. For example, if the purpose of the user evaluation is to detect how users will interact with the interface of a system – in this

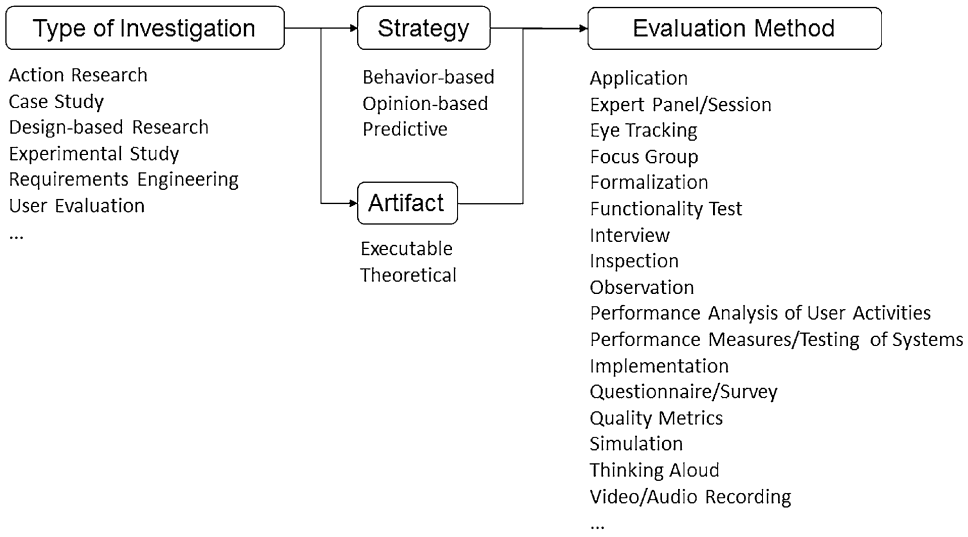

Fig. 1 Overview of the main concepts found in PAIS papers with focus on evaluation

case the interface is the artifact –, evaluation methods, such as observation or thinking aloud, are effective methods to obtain information about their behavior. In the computer science domain, it is necessary to distinguish between *executable* artifacts (e.g., (high-fidelity) prototype) and *theoretical* artifacts (i.e., non-executable artifacts) such as designs or concepts. Depending on the type of artifacts, different evaluation methods are more or less adequate. For example, methods for the analysis of performances are only relevant for executable artifacts. Based on Gediga and Hamborg (2001), the evaluation methods can be classified according to their strategic direction (see *Strategy* in Fig. 1): *behavior-based*, *opinion-based*, and *predictive*. The decision for this classification is described in Sect. 2.3. For example a behavior-based evaluation method concentrates on the analysis of users' activities during their interaction with the investigated artifacts (e.g., observation methods) whereas an opinion-based evaluation method identifies users' views on the investigated artifacts (e.g., interviews). The predictive evaluation method has its focus on the investigation of the context, feasibility as well as performances of the artifacts (e.g., functionality tests). For instance, behavior-based evaluation methods are promising for the above-mentioned example of evaluating the interface of a system in order to obtain information how users will interact with the system.

The analysis of the papers for our literature review was particularly focused on *Artifact*, *Strategy*, and *Evaluation Method*. In contrast to *Type of Investigation*, which can include a combination of several different evaluation methods, *Evaluation Method* was considered as an elementary step to evaluate PAIS artifacts. Although evaluation methods have repeatedly been summarized in Information Systems, Business Informatics, or Software Engineering, there seems to be a lack of contributions illustrating research and use of evaluation methods in the context of PAIS contributions. The goal of this article is to identify these elementary evaluation methods and to categorize them according to their strategic direction.

2.2 Related Work

As stated in the introduction, a scientific analysis of evaluation methods applied in PAIS with respect to human orientation has been lacking until now. In the past years, PAIS research has centered on artifact-related evaluation methods. For example, evaluation methods for process mining and data mining are investigated in Rozinat et al. (2007). Often evaluation methods are reviewed as part of the development of evaluation frameworks with focus on different aspects of PAIS. For instance, an evaluation framework with the goal to assess a business process modeling tool in every phase of the development process is displayed in Effinger et al. (2011). Evaluation is used as a criterion for the construction of business process reference models in Fettke et al. (2006). Lastly, a literature review by Leitner and Rinderle-Ma (2014) states that research evaluation is a challenge in the context of security in PAIS.

Furthermore, related literature can be found in the Software Engineering domain. A variety of assessments of evaluation strategies exist in Software Engineering (e.g., Gediga and Hamborg 2001; Glass et al. 2002) and for specific domains such as adaptive systems (cf. van Velsen et al. 2008) or visualization (cf. Shneiderman and Plaisant 2006; Kriglstein et al. 2014; Pohl 2012). Furthermore, Moody (2003) describes an evaluation framework with a set of metrics for assessing the quality of data models. However, none of these models center specifically on PAIS.

2.3 Classification of Evaluation Methods

As stated in the previous section, evaluation methods in PAIS have only been selectively looked into so far. A systematic analysis is missing. In order to investigate evaluation methods in PAIS, we examined classifications from related domains, specifically from Software Engineering (e.g., Gediga and Hamborg 2001; Glass et al. 2002). For example, Glass et al. (2002) classify evaluative approaches into four categories: *Deductive*, *Interpretive*, *Critical*, and *Other*. However, this approach does not meet our requirements as the *Other* category is not distinctive and cannot be easily identified. Gediga and Hamborg (2001) suggest categories based on the purpose, i.e., *Behavior-based*, *Opinion-based*, and *Predictive*. As this categorization is well suited when evaluating user behavior, authorizations or human skills, we adopted it in this paper. In addition, it does not include a category "Other" that remains undefined. Instead, it provides a categorization that is highly distinctive. The three categories of evaluation methods are described in the following (adapted from Gediga and Hamborg 2001):

- *Behavior-based* This category includes evaluation methods that collect data from users in order to analyze users' behavior during the interaction with the investigated artifacts, for instance, to identify if users interact with the artifacts in a planned manner. Representative methods are, for example, observational techniques, thinking aloud protocols and eye tracking. Moreover, behavior-based methods analyze user-centered performance, for instance log files analysis of the recorded users' interaction with artifacts or the analysis of time that the users need to solve tasks with a prototype.
- *Opinion-based* Opinion-based methods evaluate users' opinions with regard to the investigated artifacts, e.g., through questionnaires and interviews. These methods can be helpful to detect suggestions for improvements from users or to assess the satisfaction of users for investigated artifacts.
- *Predictive* Evaluation methods of this category aim at assessing the context of use of investigated artifacts depending on different requirements (e.g., domains, systems, and users). Typical examples are inspections, walkthroughs, use cases, and scenarios. For example, an inspection can be applied to analyze the investigated artifacts in regard to usability heuristics. Such evaluation methods can already be used very early in development process (e.g., to investigate which tasks the users want to perform and if these tasks could be carried out with the investigated artifacts). Moreover, these methods assess the investigated artifacts in regard to their feasibility (e.g., prototypical implementation of a concept) and to their performances in order to evaluate the artifacts under realistic conditions (e.g., execution time of an algorithm).

This categorization centers on the purpose of evaluation methods, i.e., for which aim they are applied, and can be assigned to more than one category. For example, if the aim of using a focus group is to ask a group of people about their perceptions and opinions, then this method is classified as belonging to the category *Opinion-based*. However, if the focus group method is used to analyze the investigated artifacts in order to evaluate their utility and feasibility for a group of people, then the category is *Predictive*. In the next sections, we will utilize this categorization in order to classify the investigated evaluation methods.

3 Methodology

Figure 2 displays the overarching methodology of this paper. It can be seen from the figure that a systematic literature review, an expert survey, and a focus group were conducted. The aim of this multi-method study is to identify evaluation methods not only typical of PAIS, but which also have a focus on aspects that are particularly relevant for humans working with the system (including human orientation in general, graphical presentations (visualization), and security topics since security can fail due to misunderstandings or differing underlying assumptions of users and systems). In particular, our research was guided by the following questions (RQ):

RQ1: Which evaluation methods are typically used? This question aims to identify which evaluation methods are currently utilized to assess and evaluate artifacts for human orientation in general, but also for security and visualization in PAIS.

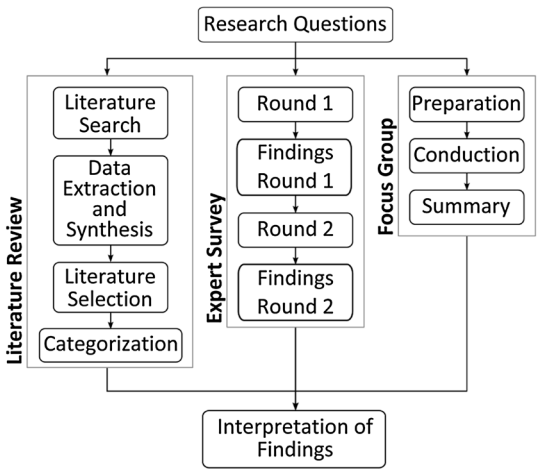

Fig. 2 Methodology

RQ2: Which evaluation methods are of future interest? The aim is to detect further methods that can be of interest for the evaluation of artifacts with focus on human orientation in general, but also on security and on visualization.
RQ3: How can these evaluation methods be classified into the categories Behavior-based, Opinion-based, and Predictive based on the purpose of evaluation? We analyze if typical and future evaluation methods can be categorized into *Behavior-based*, *Opinion-based*, and *Predictive* in order to examine the applicability of this classification.

To answer the research questions, we first conducted a literature review (cf. Cronin et al. 2008) to provide a critical analysis and overview of relevant available literature contributions in order to identify which evaluation methods have been used in the last years (see research question RQ1).

Additionally to the literature review, we conducted a two-round online survey with experts in the field of human orientation, security, and visualization in PAIS. The aim of the expert survey was to gain a closer insight into current (evaluation) practice in research, to confirm typical methods, and to identify existing evaluation methods that might not be disseminated in publications. In the first round we concentrated on questions to identify a) which evaluation methods the experts already used (see research question RQ1), and b) evaluation methods that might not be often used but could be of future interest (see RQ2). Since we asked experts from different fields, we expected to obtain a diversity of evaluation methods for PAIS which also included evaluation methods that were originated from and utilized only in one of these fields. The second round had the aim to verify the identified evaluation methods and artifacts according to the categories *Behavior-based*, *Opinion-based*, and *Predictive* (see research question RQ3).

Since the answer options may be interpreted differently by respondents and we found misunderstandings regarding meanings of questions or responses within the online expert survey, we decided to conduct a focus group (cf. Stewart et al. 2007) with experts different from the expert survey to discuss the found artifacts and evaluation methods by addressing RQ1 and RQ2 as well as the methods' categorization (RQ3). This gave us the possibility not only to verify the evaluation methods and their categorizations but also to identify further evaluation methods from the discussion with experts. The main benefit of the focus group over the online survey was the interaction between the participants and the possibility to query the participants in order to avoid misunderstandings. Furthermore, the combination of both methods allowed us to collect and analyze qualitative and quantitative data which greatly contributed to answering the research questions effectively.

4 Literature Review

A literature review (cf. Cooper 1988) was conducted of contributions referring to human orientation in general, security, and visualization in PAIS in order to identify typical evaluation methods in research (RQ1), and to categorize these methods (RQ3) by using the *Behavior-based*, *Opinion-based*, and *Predictive* classification defined in Gediga and Hamborg (2001) (see Sect. 2.3).

4.1 Procedure

This section describes the procedure of the systematic literature review based on guidelines defined in Kitchenham (2004) and Kitchenham and Charters (2007). The review can be divided into four phases, i.e., literature search, literature selection, data extraction and synthesis, and categorization, all of which are described in the following.

4.1.1 Literature Search

A literature review was performed manually by using horizontal (i.e., focused) and vertical (i.e., generic) searches in order to maximize the results regarding potential literature between 09/01/2012 and 09/30/2012. For horizontal searches, we investigated the conference proceedings of the Conference on Business Process Management (BPM), the Conference on Advanced Information Systems Engineering (CAiSE), the Conference on Cooperative Information Systems (CoopIS), the Conference on Enterprise Distributed Object Computing (EDOC), and the European Conference on Information Systems (ECIS). In addition, we examined the following journals: Business Process Management Journal, Information Systems, Data and Knowledge Engineering, Decision Support Systems, European Journal of Information Systems, and MIS Quarterly. We selected conferences and journals that are well known to the BPM community and have a low acceptance rate or a large impact factor. Furthermore, vertical searches were conducted by means of the search engine Google Scholar (http://scholar.google.com) and the publisher databases ACM Digital Library (http://dl.acm.org), SpringerLink (http://www.springerlink.com), and IEEE Xplore (http://ieeexplore.ieee.org). The vertical searches were performed to discover relevant literature that was not identified by the horizontal searches. However, similar results of the conferences and journals were returned, and only few extensions were obtained.

Table 1 List of keywords, total literature hits, and number of selected papers between 1993 and 2012

Aspects	Keywords	Total hits	Selected papers
Human orientation	Human orientation, work experience, experience, resource, allocation, capabilities, organizational model, actor, human agent, human resources, work distribution, skills, capabilities, competencies, attitudes, experience process aware information systems, and workflow systems	607	59
Security	Workflow security and business process security	670	67
Visualization	Layout algorithm for business process, process model editor, worklist visualization, Process visualization, workflow visualization, RBAC visualization as well as event logs and business process visualization	1799	151

Table 2 Assignment of evaluation methods and artifacts found in the literature review with regard to their categories

Category	Evaluation methods
Behavior-based	Thinking aloud, observation, and video/audio record analysis
Opinion-based	Questionnaire, interview, and focus group (includes group discussion)
Predictive	Application (includes case, example, scenario, storyboard, and use case), contextual inquiry method, expert panel, formalization, function tests, implementation (includes prototypical implementation), inspection (includes heuristics and reviews), simulation, and performance measures (includes measurements of the artifacts like complexity measures, precision measures, generalization measures, robustness measures, precision and recall metrics, and execution time)
Type	**Artifacts**
Executable	Algorithm, implementation, prototypical implementation, and system
Theoretical	Algorithm, architecture, concept, environment, framework, guidelines, literature, mechanism, methodology, model, pattern, requirements, strategy, and theory

4.1.2 Literature Selection

The literature was selected according to the following criteria: text availability, duplicate reduction, relevance of the title and abstract, and further analysis of relevant text segments identified by means of keyword search in the texts in order to identify evaluation methods and artifacts. For security, the publication set was additionally refined for human-centered research and hence only publications centering on humans (often called users, agents, or resources in literature) were selected. Table 1 shows the keywords, the total hits of found publications and the number of selected papers. It can be seen from Table 1 that each area has a unique set of keywords. Initially, we received a large number of total hits on potential literature. To identify the relevant literature, we reviewed the title, keywords, and content of the paper to examine if the publication included evaluation methods, if it was user-oriented, and if it could be allocated within the research on human orientation, security, or visualization in PAIS. This procedure reduced the number of selected publications to in total 277 which centered on human orientation, security, or visualization in PAIS (see Table 1).

4.1.3 Data Extraction and Synthesis

Given the resulting set of selected papers, the next step was to extract the information about theoretical and executable artifacts and the evaluation methods applied in the papers. This process is similar to a qualitative content analysis (cf. Auer-Srnka and Koeszegi 2007; Mayring 2003). Generally, it is expected that authors specifically name and describe their used artifacts and evaluation methods in the paper. However, we found that only few clearly stated the utilized evaluation methods and artifacts. This complicated the data extraction as in all other cases, we had to skim the content of the paper in order to identify and define the method. Moreover, we noted that often different terms are used in the publications for similar evaluation methods. For example, in some publications the term "example" was utilized while in others the term "scenario" was used to describe or display the artifact in an example. In our findings we therefore aggregated these terms into a category *application*. Thus in Table 2, *application* includes the terms cases, examples, scenarios, storyboards, and use cases because these artifacts were used in a similar context in the publications.

4.1.4 Categorization

The categorization was conducted utilizing the classification of *Behavior-based*, *Opinion-based*, and *Predictive* evaluation methods (see Sect. 2.3). In particular, three of the four authors discussed each evaluation method and categorized it into one of these categories. Then, the results were discussed and validated by the fourth author who performed an independent categorization.

In addition, the artifacts were classified as theoretical and executable in a similar way. The classification session led to vivid discussions on whether to categorize artifacts as theoretical or executable. For example, an algorithm can be a theoretical and an executable artifact. We also received mixed results in the expert survey (see Sect. 5) and similar discussions in our focus group (see Sect. 6). In this study, we acknowledge that these ambiguities exist, and we dealt with the problem by carefully reading of the content of each paper in order to clearly identify the artifacts. This is further analyzed in the discussion in Sect. 8.

In conclusion, the systematic literature review comprised four steps: first, an extensive literature search was conducted spanning the areas of human orientation, security, and visualization in PAIS. Second, the literature was selected by mainly analyzing the title, keywords, and content. In the next step, artifacts and evaluation methods were extracted. Lastly, the extracted artifacts and evaluation methods were categorized into *Behavior-based*, *Opinion-based*, and *Predictive* based on Gediga and Hamborg (2001).

4.2 Results of the Literature Review

Table 2 displays all evaluation methods and artifacts concerning human orientation in general, security, and visualization in PAIS. As can be seen from the table, each evaluation method is categorized in *Behavior-based*, *Opinion-based*, or *Predictive*. A complete list of the literature found regarding the type of artifacts and evaluation methods is given in the Appendix (available online via http://link.springer.com/).

In addition, Fig. 3 shows a graph visualizing the connections (edges) between the different categories of evaluation methods (orange nodes) and the investigated artifact types (green nodes) based on the findings of the literature review. Size and numbers of nodes reflect the number of found artifacts/evaluation methods. The thickness and the number of edges displays the number of connections between evaluation methods and artifacts. It must be pointed out that for some artifacts more than one method was applied. For example, Fig. 3 shows that different evaluation methods from the category *Predictive* were used as combination for the evaluation of the same artifact (e.g., for the evaluation of a prototype the application and performance measures were used as predictive evaluation methods). Table 3 presents the number of the different evaluation methods and the investigated artifacts types for human orientation in general, security, and visualization respectively in detail.

Table 3 Numbers of found artifact types and evaluation method categories in literature. It must be pointed out that some literature contributions included more than one method and artifact

Category	Human orientation	Security	Visualization	Total
Behavior-based	11	2	9	22
Opinion-based	17	4	18	39
Predictive	95	102	244	441
Executable	19	10	116	145
Theoretical	47	71	103	221

Most of the selected literature contributions applied evaluation methods from the category *Predictive* (cf. Fig. 3). Especially used were evaluation methods from this category which do not consider users in their evaluation. For example, it was observed that implementation as evaluation method was applied in order to verify the theoretical artifacts (found in 21 of 59 publications in the area of human orientation, security: 37/67, and visualization: 112/151). Furthermore, applications include cases, examples, scenarios, and use cases (found in 36/59 papers of human orientation, security: 54/67, and visualization: 79/151) and were often used for the inspection of artifacts in order to ensure that the requirements are fulfilled.

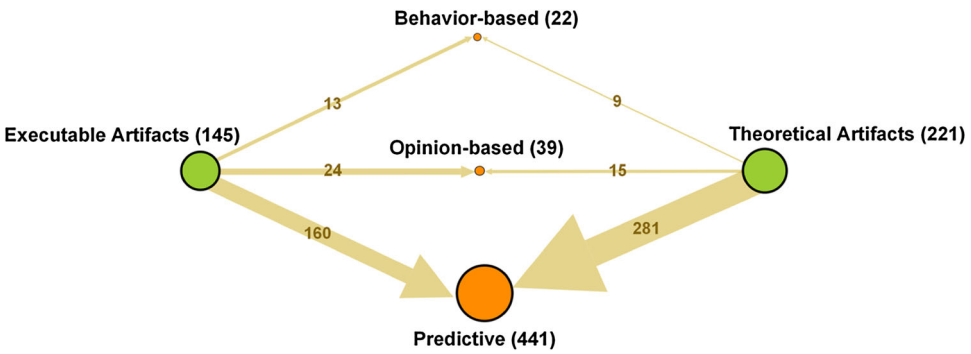

Fig. 3 Connections between the different categories of evaluation methods (*orange* nodes) and the investigated artifacts types (*green* nodes) based on the findings of the literature review. The numbers show how often the categories of methods/artifact types were mentioned and how often the categories of methods were applied to the artifacts types. It must be pointed out that for some artifacts more than one method was applied

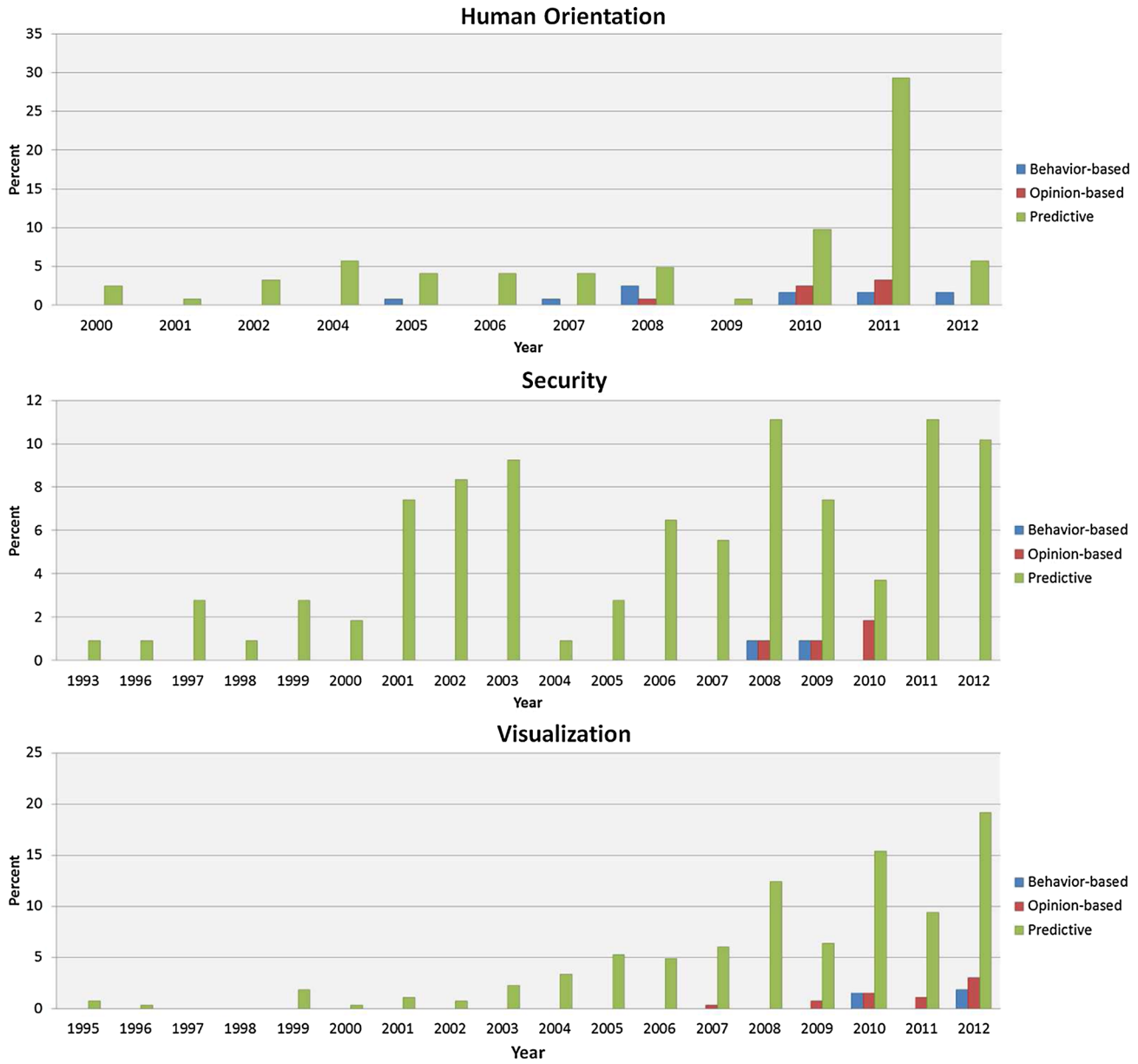

Fig. 4 Distribution of the evaluation methods found in the literature for human orientation in general, security and visualization, taking into account that a literature contribution can apply to more than one evaluation method

In comparison with the predictive evaluation methods, behavior-based and opinion-based evaluation methods were applied sparingly (human orientation: 28/59 publications, security: 6/67, and visualization: 27/151). However, a closer look shows that the number of literature contributions which include evaluation studies with users has increased in the last five years (cf. Fig. 4).

It was observed that in many cases (human orientation: 8/59 publications, security: 26/67, and visualization: 90/151), a combination of different evaluation methods was used. For all three research areas, frequently methods from the category *Predictive* were combined, e.g., prototypical implementation to prove the concept with scenarios/use cases and/or performance measures. In visualization of PAIS, combinations of behavior-based and opinion-based evaluation methods were, e.g., thinking aloud with observation and questionnaires/interviews.

5 Expert Survey

An online expert survey (cf. Courage and Baxter 2004) was used to complement the literature search and to address research questions RQ1, RQ2 and RQ3 (see

Table 4 List of typical evaluation methods and artifacts identified by the participants (RQ1)

Category	Evaluation methods
Behavior-based	Eye tracking, observation, performance analysis of user activities, and thinking aloud
Opinion-based	Interview, questionnaire, and expert session
Predictive	Card sorting, conformance checking, data sensitivity analysis, discourse analysis, (expert) inspection, focus group, heuristic evaluation/heuristics, model checking, performance measures, policy formalization, prototype (including wizard of oz), quality metrics, review, simulation, soundness, case/use case/scenario, and walkthrough
Type	**Artifacts**
Executable	Encryption algorithms, process mining algorithms, authentication, execution monitor, information system, process/runtime engine, (hi-fidelity) prototype, prototypical implementation, user interface, software mockup, and user interface/worklist
Theoretical	Access control policy, conceptual model (data/process), data to visual representation mapping, domain-specific (modeling) language (DSL/DSML) for the specification of process-related security properties, initial sketches, knowledge map, organizational model, paper mockup, mockup, paper prototype, platform-specific model (PSM) with process-related security properties, process logs, process model, process priority model, quality framework, requirements description, security requirements documents, scenarios, security ontology, task to visualization mapping, usage control policy, use cases, use case descriptions, and use cases/functional descriptions

Sect. 3). The aim of the expert survey was to gain further insights into current (evaluation) practice in research which has not necessarily been published (e.g., when papers concentrate only on the presentation of their concepts without providing an explicit description of the evaluation), to confirm typical methods and to identify evaluation methods that might not be disseminated in the publications. The usage of an online questionnaire had the advantages that (a) it was easy to use for the participants, because they could answer the questions as it suited on their schedule, (b) there was no restriction in regard to the location, and (c) the design of the survey as well as the preparation and analysis of the gained data were less costly and less time-consuming than when conducting face-to-face interviews.

5.1 Sample

The literature review provided us with a comprehensive overview of researchers who were working in the field of human orientation in general, security, and visualization in PAIS. We contacted this pool of researchers by email and asked them to take part in a survey concerning evaluation methods. They were selected because they are actively publishing in the field and together cover all three areas of human orientation, security, and visualization. In total, we received a positive response from 14 researchers (4 researchers with expertise in visualization, 6 researchers with expertise in security, and 4 researches with expertise in human orientation in PAIS). For each of these 14 participants, we created an account to provide them with access to the survey.

5.2 Procedure

The expert survey consisted of two rounds. In the first round, the participants were asked to define their level of knowledge, to name at least three typical artifacts for human orientation, security, or visualization in PAIS, to specify at least two typical evaluation methods for these artifacts, and to find prospective evaluation methods. For the second round, we used the defined artifacts and evaluation methods from the first round and categorized the evaluation methods with the previously specified classification: *Behavior-based*, *Opinion-based*, and *Predictive*. In the second round, 12 participants rated the classification of *all* evaluation methods found in the first round, specified the relevance of evaluation methods with regard to theoretical and executable artifacts, and defined missing evaluation methods that were not previously mentioned.

5.3 Results of the Expert Survey

RQ1: Table 4 displays examples of typical evaluation methods and artifacts found in the first round of the expert survey. Figure 5 shows the connection between the typical evaluation methods with regard to the theoretical and executable artifacts. It can be seen from the figure that most of the named typical evaluation methods for the theoretical artifacts belong to the category *Predictive*. This corresponds with the findings from the literature review. In contrast to the literature review, the opinion-based evaluation methods were often listed for theoretical artifacts. On the other hand, the experts named evaluation methods from the category *Predictive* and from *Behavior-based* for the executable artifacts. Security experts mentioned evaluation methods from the category *Predictive* most. Experts in human orientation named methods from the category *Opinion-based* most, and evaluation methods from the category *Behavior-based* were most stated by the visualization experts.

RQ2: Table 5 shows future evaluation methods identified by the participants. As can be seen from Table 5, prospective evaluation methods that were not mentioned as

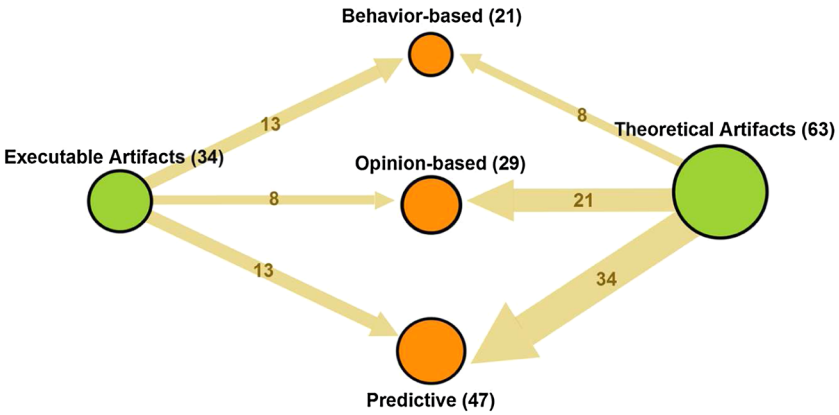

Fig. 5 Graph visualization to show the connections between the categories of typical evaluation methods (*orange nodes*) and artifacts (*green nodes*) mentioned by the experts in the expert survey. The *numbers* present how often the categories of methods/artifact types were named and how often the categories of methods were assigned to the artifacts types by the experts. It must be pointed out that for some artifacts more than one method was applied

Table 5 List of prospective evaluation methods identified by the participants. Evaluation methods that are not also named as typical evaluation methods are highlighted in boldface

Category	Evaluation methods
Behavior-based	**Emotion tracking**, eye tracking, **insight-based evaluation**, **neuroscience methods**, **neuroscientific analysis**, and observation
Opinion-based	Questionnaire
Predictive	Card sorting, **collaborative ratings**, **consistency checking**, data sensitivity, **dataflow correctness**, discourse analysis, performance measures, review, **semiotic analysis**, simulation, **user access rights evaluation**, and walkthrough

typical methods are listed in boldface (cf. Table 4). Eye tracking, questionnaire, and walkthrough were mentioned methods by all participants; each participant could mention several different evaluation methods.

Similar to the typical evaluation methods, the experts frequently mentioned evaluation methods from the category *Predictive* for theoretical artifacts, and behavior-based and predictive evaluation methods for executable artifacts. Whereas for security experts the most frequently mentioned evaluation methods were also from the category *Predictive*, experts in human orientation noted more behavior-based evaluation methods.

RQ3: In the second round, the participants supported our categorization of behavioral-based and opinion-based evaluation methods. Also in the case of predictive evaluation methods, the majority supported the classification. The focus group was the only method that was specified as opinion-based by six participants. The participants stated that focus group can include elements of the category *Predictive* but also of the category *Opinion-based*, because it can combine an inspection with a group discussion with focus on different viewpoints and opinions about artifacts. The definitions given by the participants in the first round of the expert survey already showed that their opinions were divided in regard to the focus group: "sessions with groups of users to collect information about requirements, current usage, current problems, etc." and "a focus group is built in order to evaluate the proposed knowledge map against the background of underlying human-oriented processes", while others define it as "[...] opinions on visual representations" and state that "questions are asked in an interactive group setting where participants are free to talk with other group members".

The results let us conclude that the participants were able to apply the used classification to the collected evaluation methods.

6 Focus Group

Although we gained valuable insights from the expert survey, we also identified misunderstandings about the meanings of questions or contradictory responses. Therefore, we decided to conduct a focus group session (cf. Courage and Baxter 2004; Stewart et al. 2007) which is an effective method to analyze the results from different points of view in a short period of time. The interaction between the participants and the possibility to ask the participants questions makes it possible to avoid misunderstandings and to verify the found evaluation methods and artifacts in a discussion round. In particular, the focus group session gave us the possibility to a) discuss and verify the results from the literature review (RQ1) and from the expert survey (RQ1, RQ2, and RQ3), but also to b)

collect and identify missing or further evaluation methods (RQ2) resulting from the discussion and interaction between participants in the group.

6.1 Sample

Since we decided on a face-to-face focus group (which allows the exchange of visual and nonverbal cues to enhance communication), we were restricted to inviting people from the local area. Further criteria for recruiting the participants were (a) that they were familiar with the topic, and (b) that they had the time and interest to attend a focus group session with a duration of about one hour. In addition, we selected experts who had not taken part in the expert survey in order to avoid that participants felt the need to defend the results that we gained from the expert survey. Based on the literature review and expert survey results we observed a trend toward *Behavior-based* and *Opinion-based* evaluation methods. Since these are well-known methods in Human Computer Interaction, it was important for us to have at least one participant with expertise in this field to identify further methods known in Human Computer Interaction but which have not been adopted in PAIS so far. A further criterion was that the participants covered the key concerns of business process management defined in van der Aalst (2012) and had a comprehensive knowledge of evaluation methods in computer science. Finally, we selected four senior researchers with expertise in human orientation, security, and/or visualization in the context of PAIS. One of these experts also had additional expertise in Human Computer Interaction.

6.2 Procedure

The focus group was conducted in a university meeting room and took about one hour. The session was guided by two skilled moderators and one note taker who helped the moderators. The focus group session consisted of two steps: First, the participants filled out a questionnaire in which they had to (1) define their level of knowledge, (2) grade the relevance of typical and prospective evaluation methods for theoretical and executable artifacts in PAIS found in the first round of the expert survey, and (3) find future evaluation methods. In the second step, the participants discussed the relevance of evaluation methods and possible future directions for theoretical and executable artifacts.

6.3 Results of the Focus Group

RQ1: The participants discussed the set of evaluation methods which resulted from the first round of the expert survey. They found it was an arbitrary and fuzzy set of methods and mentioned that some methods did not seem to be evaluation methods but rather artifacts (e.g., policy formalization). The reason for this difference between the two groups is that the focus group allows discussions between experts in order to reduce misunderstandings between the individual interpretations which is not feasible in a survey. Moreover, the participants agreed that the users should play a more important role in evaluation methods especially for executable artifacts. As users work with the system on many levels, e.g., designers and analysts, they should be more involved in the evaluation.

RQ2: Although the experts found it difficult to define future evaluation methods at the beginning and only interdisciplinary evaluation methods were mentioned, e.g., PAIS research in connection with social scientific methods, the discussion led them to detect missing evaluation methods which might also be of interest for the future: qualitative comparison with existing systems/prototype, case study, evaluation along identified threats, log file analysis, statistical evaluation, experiments, logging machine behavior, granularity analysis, and ethnography/grounded theory.

RQ3: The participants stated that for rating the relevance additional information, such as which artifacts the methods relate to or a category scheme (e.g., theoretical, technical, and human related evaluation), is missing. A classification of methods would further support the relevance rating of evaluation methods (in the questionnaire). We proposed our category scheme (*Behavior-based*, *Opinion-based*, and *Predictive*), and the participants agreed to it.

7 Summary of Evaluation Methods

This section summarizes and outlines the results of the literature review (see Sect. 4), expert survey (see Sect. 5), and focus group (see Sect. 6).

An extensive list of evaluation methods for the area of human orientation in general, security, and visualization in PAIS are shown in Table 6 which displays a classification of the typically used evaluation methods identified by the literature review, expert survey, and focus group. The table shows which evaluation methods belong to which category: *Behavior-based* (*B*), *Opinion-based* (*O*), or *Predictive* (*P*). These methods are used in various ways for theoretical and executable artifacts. Hence, the artifact column displays whether the evaluation method is examining a *Theoretical* (*T*) or *Executable* (*E*) artifact. For instance, interviews were conducted for theoretical as well as executable artifacts. The last three columns indicate in which area, *Human orientation* (*Hum*), *Security* (*Sec*), or *Visualization* (*Vis*) the method was utilized or indicated. Please note that this list of evaluation methods reflects the

Table 6 Summary of evaluation methods; described by name, category (behavior-based (B), opinion-based (O), and predictive (P)), artifacts (executable (E) and theoretical (T)), and areas (human orientation (Hum), security (Sec), and visualization (Vis))

Evaluation methods	Category	Artifact	Hum	Sec	Vis
Application	P	E	☒	☒	☒
(includes Case, Example, Scenario,		T	☒	☒	☒
Storyboard, and Use Case)					
Card sorting	P	T	☒	☐	☐
Contextual inquiry	B	T	☒	☐	☐
Conformance checking	P	E	☐	☒	☐
Data sensitivity analysis	P	T	☐	☒	☐
Discourse analysis	P	T	☐	☒	☐
Correctness	P	T	☒	☒	☐
(includes formalization, model checking,					
and Soundness)					
Expert panel/session	O	E	☒	☐	☐
		T	☒	☐	☐
Eye tracking	B	E	☒	☐	☒
		T	☒	☐	☒
Focus group	O	E	☐	☐	☒
(includes Group Discussion)		T	☒	☒	☒
	P	T	☐	☐	☒
Functionality test	P	T	☒	☐	☐
Implementation	P	T	☒	☒	☒
(includes prototype)					
Interview	O	E	☒	☒	☒
		T	☒	☒	☒
Inspection	P	E	☒	☐	☒
(includes heuristics and review)		T	☒	☒	☒
Observation	B	E	☒	☒	☒
		T	☒	☒	☒
Performance analysis of user activities	B	E	☒	☐	☒
Performance measures/testing of systems	P	E	☒	☒	☒
Questionnaire	O	E	☒	☒	☒
		T	☐	☒	☒
Quality metrics	P	T	☐	☒	☐
Simulation	P	E	☒	☒	☐
		T	☒	☐	☐
Thinking aloud	B	E	☐	☐	☒
		T	☒	☒	☒
Video/audio recording	B	E	☒	☐	☒
		T	☒	☐	☐
Walkthrough	P	E	☒	☒	☒
		T	☒	☒	☒

results from the literature review, expert survey, and focus group. The list does not claim to be exhaustive and can be extended in further studies.

The analysis of evaluation methods showed us that different words were used for the same kind of evaluation methods (e.g., example or application was used synonymously for use case). Hence, for a better readability, we grouped evaluation methods which were used in the same context, as can be seen in Table 6. For example, the *application* method includes also case, example, scenario, storyboard, and use case methods. All these methods were used in the reviewed literature and were also stated by the experts in the expert survey and focus group to describe the application of an artifact, e.g., by using task descriptions or test cases. Furthermore, discussion in groups and *focus group* as well as *implementation* and prototype as methods

were similarly used in the literature but also by the experts. The *inspection* method includes review techniques to detect a large number of basic problems, considering, e.g., a set of guidelines, heuristics, or standards. Therefore, the *inspection* method incorporates also the review and heuristic methods.

To sum up, we identified a set of 23 evaluation methods. Even though we identified some evaluation methods suitable for only certain areas (such as functionality tests in human orientation), most of them can be adapted and applied to other areas such as security and visualization.

8 Discussion

Based on the previous section, we will discuss results, recommendations, lessons learned, the potential impact on research and practice as well as limitations of this paper.

8.1 Results

In this article, we investigated and examined artifacts as well as evaluation methods in the area of human orientation in general, security, and visualization in PAIS. A complete list is shown in Table 6 in the previous section. This list can be used as reference to evaluate theoretical and executable artifacts. In the following, we will describe four results derived from the literature review, expert survey, and focus group.

8.1.1 Result 1: Focus on Predictive Evaluation Methods

In the literature review, we noted that behavior-based and opinion-based evaluation methods are less frequently used than predictive evaluation methods. We assume that during the past 30 years, PAIS research has centered on the design and development of core PAIS-relevant features such as implementation, function, and application. Behavior-based and opinion-based methods focus on users activities and feedback. It can be seen from the literature that the use of these methods has not been the main focus so far. Based on these results, we can assume that the technical quality of PAIS has improved while user experience and feedback have been neglected during the development. However, user evaluations conducted in the past PAIS developments might not have been published.

8.1.2 Result 2: Ten Widely used Evaluation Methods

From 23 evaluation methods in Table 6, the following 10 evaluation methods were applied in all three areas:

- performance measures/testing of systems and questionnaires for executable artifacts
- implementations, inspections, focus groups, and thinking aloud for theoretical artifacts
- applications, interviews, observations, and walkthroughs for theoretical as well as for executable artifacts

Five of ten evaluation methods that are used in all three areas are predictive evaluation methods, followed by three opinion-based and two behavior-based evaluation methods. This also reflects the previously found prevalence of predictive evaluation methods.

8.1.3 Result 3: Interdisciplinary Evaluation Methods

In the following, we will highlight the evaluation methods that do not belong to the 10 widely used evaluation methods and can be found in only two of the areas: human orientation, security, and/or visualization.

Human Orientation and Security Correctness evaluations of theoretical artifacts and simulations for executable artifacts were mentioned by the experts and in the literature as evaluation methods for the areas human orientation and security. Both methods are also interesting methods for the area of visualization, e.g., to simulate different visualization layouts or to verify the correctness of a layout algorithm.

Human Orientation and Visualization Eye tracking, performance analysis of user activities, and video/audio recording only appeared in the areas human orientation and visualization. These three methods are all behavior-based methods and are primarily used for executable artifacts. The literature review showed that for security, the evaluation of theoretical artifacts played a more important role in the last years than the evaluation of executable artifacts. Nevertheless, these three methods can also be applied for the evaluation of security relevant executable artifacts if the analysis of users' behavior is of interest.

Security and Visualization The results showed that except for the ten widely used evaluation methods, no explicit, distinct, and overlapping evaluation methods between the areas of security and visualization were discovered. Only questionnaires for theoretical artifacts were found. However, we also identified questionnaires for executable artifacts in the human orientation area. This does not mean that no existing methods exist which intersect the areas security and visualization. In this study based on the results of the literature review, expert survey, and focus group, however, we were not able to identify them.

8.1.4 Result 4: Trend Towards Human Orientation

In recent years, research has studied and analyzed users in PAIS more frequently (e.g., Mendling et al. 2007; Mendling and Strembeck 2008; La Rosa et al. 2007; Kabicher-Fuchs et al. 2012). This tendency is also reflected in the results of the expert survey and focus group. Although findings of the literature review showed a larger gap between predictive and opinion-based/behavior-based methods, the results of the expert survey and focus group highlighted an increasing usage of opinion-based and behavior-based methods. This increased interest in human aspects of PAIS is also reflected in current conference calls such as highlighted in the introduction section (see Sect. 1).

8.2 Recommendations

Based on the results, we propose the following three general recommendations. These recommendations provide researchers with an overview of aspects which they should consider in their investigations.

8.2.1 Recommendation 1: Choose Evaluation Methods based on Research Goals

The selection of an evaluation method depends strongly on the objectives that your work is aiming at. Artifacts, data type, time, feasibility, and monetary funds are essential factors to consider when choosing the adequate evaluation methods. For example, a walkthrough can be used to analyze usability issues in software products.

8.2.2 Recommendation 2: Use a Mix of Evaluation Methods

As can be seen from Table 6, a large amount of evaluation methods exists for validating and testing research artifacts. Not only more common evaluation methods but also selectively used ones can be utilized. For example, the methods card sorting, contextual inquiry, expert panel/session, and functionality test were only found for the area of human orientation. However, expert panels and functionality tests can be used in all three areas, e.g., to discuss security-related topics with experts or for testing the functionality provided by a visualization system. Furthermore in the area of security, the evaluation method discourse analysis aims to investigate socio-psychological characteristics of individuals and can also be of interest for the area human orientation to, for instance, find out more about people's work experience.

8.2.3 Recommendation 3: Clearly Indicate Artifacts and Evaluation Methods in Publications

This might be surprising but during our review of literature, the artifacts and evaluation methods were often not explicitly stated. We recommend that authors provide a full description of their evaluation methods. Examples of artifacts and evaluation methods can be found in Tables 2 and 4. This can facilitate the reading of publications and promote systematic reviews on evaluation methods.

8.3 Lessons Learned

We observed that experts in the expert survey and in the focus group referred to evaluation methods on different abstraction levels (e.g., review versus model checking), to different theories, and types of evaluation (e.g., grounded theory or usability evaluation). A possible reason is that often it is difficult to draw a clear boundary between the different granularity of evaluation methods.

Furthermore, multiple definitions of evaluation methods exist in human orientation, security, and visualization in PAIS. For example, case studies were mentioned as evaluation methods by experts in the expert survey and in the focus group. However, according to Yin (2003), a case study is a strategy and includes methods like interviews and participant observation for data collection. Hence, a framework specifically for PAIS that describes different evaluation strategies including artifacts and evaluation methods would be helpful as a common basis.

A reason for the different definitions and interpretations between experts could be that the participants came from different domains (e.g., human orientation, security and visualization in PAIS). Nevertheless, this diversity of experts had the benefit of collecting typical evaluation methods for PAIS from different viewpoints (e.g., systems engineering methods, human computer interaction methods, and social scientific methods). Moreover, the usage of interdisciplinary evaluation methods was perceived as gaining importance for future research.

However, not only the definitions of evaluation methods varied but also the meaning and its context differed between the fields human orientation, security, and visualization. For example, in process mining (cf. van der Aalst 2011) a log file analysis typically consists of the examination of (process) event logs which represent process execution histories. However, in the Human Computer Interaction domain the users' activities (e.g., mouse clicks) are logged. Therefore, a taxonomy to provide a common understanding and contextual meaning would support the understanding of common practices and should be

combined with the above mentioned framework for the different evaluation strategies.

8.4 Limitations

In the literature review, the classification of the publications was performed based on (1) the content of the publication, and (2) the textual definition. By analyzing the content, we ensured that the misuse of definitions (e.g., a use case instead of a scenario) would not alter our results. During the review, we discovered that the studies are often not fully described in publications. For this reason, we skimmed the text headings and captions of figures to identify artifacts and evaluation methods used in the publications. Often, we had to fully read the paper. Furthermore, we assessed the main idea behind each publication and identified artifacts and evaluation methods based on the course of actions.

Furthermore, the assignment of the artifacts to be theoretical or executable artifacts as, e.g., shown in Table 2 was vividly discussed among the authors. We noticed that in our study the experts specified an algorithm as theoretical and as executable artifact. Hence, in Table 2, an algorithm is assigned to a theoretical and an executable artifact. Another example is the prototype. Most publications use various names for this such as prototype, (prototypical) implementation or proof of concept. In visualization, a prototype may also refer to a paper mockup as a theoretical artifact. However, in human orientation and security, a prototype always refers to an executable artifact. We acknowledge that these ambiguities exist in research. Here, we dealt with this challenge by carefully reading each publication to determine which method was actually used.

In this study, it was not possible to identify which of the evaluation methods named by the experts are more or less relevant. One reason could be that the choice of evaluation method depends strongly on which artifact is going to evaluated and on the aim of the evaluation. For example, if the aim of the evaluation is to find out how users interact with the system, the information about the time a user needs to complete predefined tasks might not give enough insight into user behavior. But a combination of logging users' activities with the system, observation, and thinking aloud may be more suitable to assess users' behavior. This means that the usefulness and applicability of each evaluation method depends on the investigated artifact.

Since the number of options and applications was extremely large it was not possible to generalize the results. However, an evaluation of the different evaluation methods in regard to their specific application (i.e., theoretical and executable artifacts) in different situations is essential for research in PAIS and thus subject to future work. In order to minimize the different options, a further possible direction for future work is to concentrate on a single category of the evaluation methods and compare these methods by means of experiments.

8.5 Impact on Research and Practice

The aim of this paper was to assess how research conducts the evaluation of theoretical and executable artifacts for human orientation in general, in security, and in visualization in PAIS. For this purpose, we provided a list of these artifacts and which evaluation methods are typically used to conduct an evaluation. This collection of artifacts and evaluation methods may serve as a basis for researchers and practitioners who wish to investigate, e.g., theoretical artifacts when selecting typical evaluation methods. Furthermore, researchers and practitioners may use this collection to discover unfamiliar, interdisciplinary evaluation methods. For example in the area of security, research has neglected the evaluation of security modeling extensions (cf. Leitner et al. 2013). This paper may call attention to alternatives of how to evaluate users' preferences or understanding (e.g., observation and interview as evaluation methods).

In addition, the classification of evaluation methods can be used as a guideline for categorizing the evaluation methods researchers utilize. This paper provides an allocation of artifacts and evaluation methods in PAIS. This may serve as a basis and can be extended by adding new evaluation methods and artifacts that are not listed in this paper.

Furthermore, practitioners may use this paper for reassessing evaluation methods and to become acquainted with unfamiliar evaluation methods. This might lead to an improvement of software, for example, by using new evaluation methods in the software development.

9 Conclusion

This paper has analyzed and examined evaluation methods for human orientation in general, for security, and for visualization in PAIS. First, we conducted a literature review to assess typical evaluation methods and classified them as *Behavior-based*, *Opinion-based*, and *Predictive*. Second, an expert survey was carried out that consisted of two rounds. In the first round, the participants identified typical and future evaluation methods in PAIS. We categorized the evaluation methods found in the first round and asked the participants to rate the categorization in the second round. Third, we conducted a focus group to examine the evaluation methods found in the first round of the expert survey and to discuss future and lacking

evaluation methods which were neither named by the participants in the expert survey nor mentioned in the papers that we found. Based on the literature review, expert survey, and focus group, we summarize the main findings for each research question:

1. We discovered ten evaluation methods that are utilized in human orientation, security, and visualization in PAIS: performance measures/testing of systems, questionnaires, implementation, inspection, focus group, thinking aloud, application, interview, observation, and walkthrough.
2. The results showed that behavior-based and opinion-based methods were recognized as prospectively relevant. Furthermore, predictive evaluation methods will continue to be of importance.
3. The categorization of evaluation methods of PAIS research in the fields of human orientation in general, of security, and of visualization could be used for assigning collected evaluation methods by participants in the expert survey and the focus group.

For future work, we plan to further investigate evaluation methods. Based on the results, we will establish an evaluation framework with specified input (e.g., what requirements are necessary to perform an evaluation) and output parameters (e.g., what is the aim of the evaluation). Furthermore, we plan to combine the framework with a taxonomy for the different evaluation methods. A further interesting point for future work is to investigate the different evaluation methods in regard to their applicability and to integrate these findings into the evaluation framework.

Acknowledgments Simone Kriglstein was supported by CVAST (funded by the Austrian Federal Ministry of Science, Research, and Economy in the exceptional Laura Bassi Centres of Excellence initiative, project nr: 822746).

References

Auer-Srnka KJ, Koeszegi S (2007) From words to numbers: how to transform qualitative data into meaningful quantitative results. Schmalenbach, Bus Rev 59

Cooper H (1988) Organizing knowledge syntheses: a taxonomy of literature reviews. Knowl Technol Policy 1:104–126

Courage C, Baxter K (2004) Understanding your users: a practical guide to user requirements methods, tools, and techniques. Morgan Kaufmann, San Francisco

Cronin P, Ryan F, Coughlan M (2008) Undertaking a literature review: a step-by-step approach. Br J Nurs 17(1):38–43

Dumas M, van der Aalst WMP, ter Hofstede AH (2005) Process-aware information systems: bridging people and software through process technology. Wiley, Hoboken

Effinger P, Seiz S, Jogsch N (2011) Evaluating single features in usability tests for modeling tools. In: Proceedings of the 3rd Workshop "Methodische Entwicklung von Modellierungswerkzeugen" at Informatik 2011, GI-Edition (LNI)

Fettke P, Loos P, Zwicker J (2006) Business process reference models: survey and classification. In: Bussler CJ, Haller A (eds) Business process management workshops. LNCS, Springer, Heidelberg, pp 469–483

Gediga G, Hamborg KC (2001) Evaluation of software systems. Encycl Comput Sci Technol 45

Glass R, Vessey I, Ramesh V (2002) Research in software engineering: an analysis of the literature. Inf Softw Technol 44(8):491–506

Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. MIS Q 28(1):75–105

Houy C, Fettke P, Loos P (2010) Empirical research in business process management – analysis of an emerging field of research. Bus Process Manag J 16(4):619–661

Irani Z (2002) Information systems evaluation: navigating through the problem domain. Inf Manag 40(1):11–24

Kabicher S, Kriglstein S, Rinderle-Ma S (2011) Visual change tracking for business process models. In: Proceedings of the 30th international conference on conceptual modeling, Springer, Heidelberg, ER'11, pp 504–513

Kabicher-Fuchs S, Rinderle-Ma S (2012) Work experience in PAIS: concepts, measurements and potentials. Proceedings 24th international conference on advanced information systems engineering (CAiSE). Springer, LNCS, Heidelberg, pp 678–694

Kabicher-Fuchs S, Rinderle-Ma S, Recker J, Indulska M, Charoy F, Christiaanse R, Dunkl R, Grambow G, Kolb J, Leopold H, Mendling J (2012) Human-centric process-aware information systems (HC-PAIS). Research Report arXiv:1211.4986 [cs.HC]

Kabicher-Fuchs S, Mangler J, Rinderle-Ma S (2013) Experience breeding in process-aware information systems. In: Int'l Conf. on Advanced information systems engineering, Springer, Heidelberg, pp 594–609

Kitchenham B (2004) Procedures for performing systematic reviews. Joint technical report, Department of Computer Science, Keele University and Empirical Software Engineering, National ICT Australia Ltd

Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report EBSE-2007-01 Version 2.3, School of Computer Science and Mathematics, Keele University and Department of Computer Science, University of Durham

Klein M, Dellarocas C, Bernstein A (2000) Introduction to the special issue on adaptive workflow systems. Comput Support Coop Work 9:265–267

Kriglstein S, Pohl M, Suchy N, Gärtner J, Gschwandtner T, Miksch S (2014) Experiences and challenges with evaluation methods in practice: A case study. In: Proceedings of the fifth workshop on beyond time and errors: novel evaluation methods for visualization, ACM, BELIV '14, pp 118–125

La Rosa M, Dumas M, ter Hofstede AH, Mendling J, Gottschalk F (2007) Beyond control-flow: extending business process configuration to resources and objects. Unpublished

Leitner M, Rinderle-Ma S (2014) A systematic review on security in process-aware information systems constitution, challenges, and future directions. Inf Softw Technol 56:273–293. doi:10.1016/j.infsof.2013.12.004

Leitner M, Miller M, Rinderle-Ma S (2013) An analysis and evaluation of security aspects in the business process model

and notation. Proceedings of the 8th international conference on availability. reliability and security (ARES), IEEE, pp 262–267
Lenz R, Reichert M (2007) IT support for healthcare processes – premises, challenges, perspectives. Data Knowl Eng 61(1):39–58
Mayring P (2003) Qualitative inhaltsanalyse. Beltz, Weinheim
Mendling J, Strembeck M (2008) Influence factors of understanding business process models. In: Abramowicz W, Fensel D (eds) Business information systems, no. 7 in lecture notes in business information processing, Springer, Heidelberg, pp 142–153
Mendling J, Reijers HA, Cardoso J (2007) What makes process models understandable? In: Alonso G, Dadam P, Rosemann M (eds) Business process management, no. 4714 in lecture notes in computer science, Springer, Heidelberg, pp 48–63
Moody DL (2003) Measuring the quality of data models: an empirical evaluation of the use of quality metrics in practice. In: Proceedings of the 11th European conference on information systems, pp 1337–1352
Pohl M (2012) Methodologies for the analysis of usage patterns in information visualization. In: Proceedings of the BELIV'12 workshop: beyond time and errors – novel evaluation methods for visualization
Ramesh V, Glass RL, Vessey I (2004) Research in computer science: an empirical study. J Syst Softw 70(12):165–176
Rozinat A, de Medeiros AA, Günther C, Weijters A, van der Aalst W (2007) Towards an evaluation framework for process mining algorithms. BPM Center Report BPM-07-06, BPMcenter.org
Schulte S, Schuller D, Steinmetz R, Abels S (2012) Plug-and-play virtual factories. IEEE Internet Comput 5:78–82
Serafeimidis V, Smithson S (2003) Information systems evaluation as an organizational institution – experience from a case study. Inf Syst J 13(3):251–274
Shneiderman B, Plaisant C (2006) Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In: Proceedings of the BELIV'06 Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization, ACM, pp 1–7
Song X, Letch N (2012) Research on IT/IS evaluation: a 25 year review. Electron J Inf Syst Eval (EJISE) 15(3):276–287
Stewart DW, Shamdasani PN, Rook DW (2007) Focus groups: theory and practice. Sage, Thousand Oaks
van der Aalst WMP (2011) Process mining: discovery, conformance and enhancement of business processes. Springer, Heidelberg
van der Aalst WMP (2012) A decade of business process management conferences: personal reflections on a developing discipline. In: Barros A, Gal A, Kindler E (eds) Business process management. LNCS, Springer, Heidelberg, pp 1–16
van der Aalst WMP (2013) Business process management – a comprehensive survey. ISRN Softw Eng, article ID 507984
Vanderfeesten I, Reijers HA (2005) A human-oriented tuning of workflow management systems. Proceedings of the 3rd international conference on business process management (bpm). Springer, Heidelberg, pp 80–95
van Velsen L, van der Geest T, Klaassen R, Steehouder M (2008) User-centered evaluation of adaptive and adaptable systems: a literature review. Knowl Eng Rev 23(3):261–281
Wilde T, Hess T (2007) Forschungsmethoden der Wirtschaftsinformatik. WIRTSCHAFTSINFORMATIK 49(4):280–287
Yin RK (2003) Case study research: design and methods. Sage, Thousand Oaks