Artifact Sampling: Using Multiple Information Technology Artifacts to Increase Research Rigor

Roman Lukyanenko
University of Saskatchewan
lukyanenko@edwards.usask.ca

Binny M. Samuel
University of Cincinnati
samuelby@uc.edu

Jeffrey Parsons
Memorial University of Newfoundland
jeffreyp@mun.ca

Abstract

Researchers in many scientific disciplines routinely conceptualize information technologies (IT) as antecedents or outcomes in theoretical models. The ongoing theorizing of IT leads to a novel methodological challenge termed instantiation validity (IV). In this paper, we contribute to research on remediating IV challenges by proposing and advocating the methodological practice of artifact sampling, whereby multiple artifacts are sampled from the population of all possible artifacts (the instantiation space). Artifact sampling extends the practice of employing multiple research subjects or survey respondents, routinely used in social sciences, into the IT artifact design space. Artifact sampling is an important methodological practice that stands to increase the rigor of research dealing with software artifacts. As it is currently not being adequately undertaken in the aforementioned research, many studies may result in biased or unjustified conclusions.

1. Introduction

Many scientific disciplines routinely conceptualize design features of information technology (IT) as antecedents or outcomes in theoretical models. Often, a researcher is interested in evaluating a theory in which the IT artifact is conceptualized as a variable (e.g., in Information Systems (IS) behavioral research), or as the concrete realization of a design principle (e.g., IS design science research). We broadly refer to both types of artifact-based work as ***information technology design research (ITDR).***

To illustrate, in a typical ITDR "behavioral" project researchers may posit that creating personalized recommender systems leads to the adoption of these systems by online users due to the propensity of personalized technologies to engender trust with users (e.g., see [1]–[3]). To evaluate this theory, researchers engage in design work to select or build one or more IT artifacts that correspond to various levels of personalization. These artifacts are then used by research participants, who report their perceptions of the artifact to the researchers. These responses are then used to test the underlying theory of personalized technologies. In such a scenario, the research findings and conclusions in such a study depend in part on the design decisions taken when operationalizing the IT artifacts (i.e., during the design of the artifact itself).

ITDR is widespread in the IS, computer science, and software engineering disciplines; it is also growing in prominence in social and natural sciences. For example, an active area of research in biology, geography, astronomy, and ethnography is digital citizen science [4]–[7], where researchers seek to engage ordinary people in scientific research with the use of mobile apps and highly interactive websites that allow users to submit observations of phenomena such as wildlife, galaxies, geographic features, or cultural objects [8]–[11]. To ensure these contributions are of high quality and the systems used to capture them are intuitive and easy to use, researchers in natural and social sciences increasingly engage in the theorizing of, and experimentation with, IT. This has resulted in an overall growth of ITDR across many scientific disciplines.

However, the ongoing theorizing of IT has resulted in methodological challenges [12]–[17]. When instantiating a particular theoretical construct, there are virtually unlimited ways to operationalize (i.e., design) the feature in the corresponding IT artifact, but no clear guidance for choosing the most appropriate one. Further, while a researcher may be interested in only one particular construct (e.g., personalization), the artifact that instantiates that construct often has to include a variety of features (e.g., navigation/help buttons) to provide basic functionality and usability. These features are not chosen based on instantiating the construct of interest, but may interact with this construct in unpredictable ways, potentially affecting results and diminishing internal validity.

These concerns have resulted in a proposal for a new kind of research validity [18] – ***instantiation***

URI: http://hdl.handle.net/10125/49920
ISBN: 978-0-9981331-1-9

validity (IV) – defined as the extent to which inferences and conclusions are warranted from observations of features of IT as instantiations of theoretical constructs or design principles [19].

Instantiation validity is made of ***inner*** instantiation validity and ***outer*** instantiation validity. Inner IV or ***operationalization validity*** is the faithfulness of the operationalization of a theory or design principle into an IT artifact. If a study misrepresents a theory through a wrongly chosen artifact, the results would not apply to the underlying theory. For example, a study of relational databases would not have operationalization or inner IV if the actual database used was a NoSQL one.

Assuming a valid inner IV, the outer IV or ***conclusion validity*** concerns the extent to which conclusions are valid from a study of IT artifacts. Having operationalization validity does not guarantee conclusion validity. Outer IV takes into the account all evidence presented and the analysis undertaken in the study. For example, a study may contain multiple pieces of empirical evidence (e.g., an experiment and a case study involving IT artifacts) – each with unique inner IV concerns. Reaching an appropriate conclusion in such study involves outer IV. The aim of good ITDR scholarship is to establish and demonstrate both outer and inner IV.

Prior IS researchers have voiced concerns related to IV, albeit without using its terminology. For example, consider Iivari [20]'s conjecture that even the seemingly versatile Technology Acceptance Model (TAM) "is valid only in the cases of some IT applications" (p. 44). We fully support this claim on the grounds of IV. The original TAM model does not provide explicit guidance on how to design IT artifacts to test this theory. Incidentally, Iivari is silent on to which IT applications TAM may not be valid, possibly due to the lack of design-level specificity in TAM itself. A wrongly chosen (designed) artifact for a study may result in erroneous acceptance or rejection of one's theory.

IV concerns are part of a broader effort to reduce confounds of IS studies. Previous research in IS and related disciplines (e.g., management) raised concerns akin to IV, but the focus thus far has been on confounds resulting from the potentially unpredictable nature of the *context* in which IS development and use occurs. Researchers have warned that conclusions drawn from "idealized" scenarios in many studies may not hold for real IS development where contextual sociotechnical variables may intervene in unpredictable ways [21], [22]. As Johns cautions, "context can have both subtle and powerful effects on research results" [23, pp. 358–387]. One possible remedy that has been suggested is constructing contextualized theories with greater sensitivity to specific localized phenomena [24]–[26].

IV extends the concerns about threats to conclusions in IS studies by shifting the focus from organizational and other extraneous factors to the artifact itself. IV becomes an issue during empirical work with IT artifacts (e.g., as part of an experiment or a case study). While this issue is general, it is especially serious when the IT artifact is a functional software system (e.g., recommendation agent, mobile app) with many interacting components, as opposed to, for example, simple algorithms, conceptual modeling diagrams or isolated components (although, IV issues are present in these simpler artifacts too, see [14], [16]). The complexity of the IT artifact may prevent a researcher from using theory to fully specify how to design the artifact and how the artifact is going to behave and interact with other factors.

Instantiation validity has roots in IS design science research (DSR) [27]–[31]. Indeed, IV concerns are present when DSR artifacts are evaluated for utility [32], [33]. As part of this work, researchers seek to construct an artifact as faithful as possible to the design principle; once the artifact is constructed, researchers evaluate it to demonstrate the utility of the underlying design principles [27],[29]. While IV is a recent notion, the DSR community has been actively exploring methods and techniques for evaluating IT artifacts [27],[32],[33]. Many notions and techniques employed when evaluating IT artifacts, may be used to address the question of whether an artifact is a faithful instantiation of a design principle (e.g., for example, by tracing features of the artifact from statements in the underlying theory [36]).

In contrast to DSR, IV is more troublesome for behavioral or so called "theory-with-practical-implications" research [20, p. 40]. In contrast to DSR, the latter tends to black-box the IT artifact [15], and thus is less likely to be cognizant of, notice, and mitigate the confounds due to the complex nature of IT.

We contribute to research on IV by proposing a novel methodological practice of using multiple artifacts – which we call ***artifact sampling*** – to complement existing ways to establish the validity of artifacts. There is no definitive solution to the problem of instantiation validity (for discussion, see [37]). As Iivari [20] notes, it is generally impossible to derive specific design guidance from more general (e.g., kernel or design) theories. Other studies support the same conclusion [36],[37]. Rather than seeing existing approaches as limiting, we position artifact sampling as a complementary methodological practice that can be pursued in conjunction with other approaches.

In this paper, we consider the precursors of artifact sampling from sampling theory and stimuli sampling research in psychology and sociology. We then develop a preliminary artifact sampling method which we illustrate using hypothetical examples. We conclude the paper by outlining directions for future research.

2. From Stimuli to Artifact Sampling

We propose that one way to address the threats to inner and outer IV is by increasing the variations of the artifacts, analogous to the way researchers routinely increase the number of human participants to reduce sampling error or increase the number of questionnaire items to improve reliability. Such an approach is proposed as methodological guidance during the design process.

Sampling theory underlies much scientific experimental work [40]. Fundamental to the theory is the principle that one may generalize the results of observations only to those subjects or objects that have been sampled [41]. As early as 1940s, however, researchers pointed out a peculiar "double standard" [41], [42]. Researchers were quite eager to apply sampling theory to *subjects* (e.g., human participants, survey respondents), but almost never extended this principle to research *objects* (i.e., experimental stimuli) [43]. Even more concerning, Brunswik argued, is that over time, researchers developed a variety of systematic approaches to increase rigor in subject sampling, including statistical methods to determine sample sizes, estimate errors and biases and draw statistical inferences. Thus, seeking large sample sizes offers an ability to eliminate potentially idiosyncratic effects of differences among individual subjects [18], [44]. The theoretical premise is that the differences are assumed to be independent of: 1) any treatment effect, 2) each other, and (3) and across subjects. Therefore, the subject differences "cancel each other out" in a sufficiently large sample. In the meantime, little attention has been paid to research objects. As early as in 1943, Brunswik [43] introduced the notion of *representative designs* which argues that sampling theory *equally concerns* subjects and objects of research. Yet, the recognition of this idea has been slow. Among key objections to Brunswik's [43] argument was the effort involved in sampling objects – an argument that persists (see, e.g., [45]).

Recently, the idea of having multiple objects within treatment and control conditions has been gaining acceptance in psychology. Echoing the instantiation validity concerns described earlier, psychologists argue and show experimentally that it is generally impossible to construct ecologically valid objects such that every feature is accounted for theoretically, and that it is difficult for researchers to adequately (i.e., fully) represent and generalize to a population of objects from a single object [41], [44], [46]–[51]. This appears to be the case both for complex objects (e.g., humans – often used to instantiate independent variables in social psychology, see [51]) and simpler objects (e.g., line drawings, see [50]) commonly used in cognitive psychology. Even when the objects are quite simple (i.e., have few features and potential interactions between them), Fontenelle et al. [48] conclude: "when it is the intention of an experimenter to generalize results beyond the particular sample of objects employed, the statistical treatment of objects as a fixed effect is generally inappropriate. Thus, unless a researcher is willing to limit the generalizability of his or her findings *severely*, the effect of stimulus sampling must be considered both in the design of the experiment and in the analysis of the results." (p. 106, emphasis added).

While the benefits of involving multiple subjects in experiments and surveys have been widely recognized, the second part of the original representative design notion that suggested to do the same for objects have been neglected in experimental research. Wells and Windschitl [51, p. 1115] consider this neglect "a serious problem that plagues a surprising number of experiments," casting doubts on the validity of conclusions drawn from such studies. To increase the validity of experimental studies, more and more researchers call for ***stimuli sampling*** – selecting objects at random from the theoretical feature space [48], [51].

Sampling from a design space also occurs in the construction and validation of surveys instruments for psychometric research in IS. Straub [52, p. 150], citing Cronbach [53], notes that "an instrument valid in content is one that has drawn representative questions from a universal pool". Similarly, we propose that an artifact that is valid in content with respect to a construct is one that has features drawn in a representative way from a universal pool (of possible features that might instantiate the construct in an artifact). Straub further suggests that "a content-valid instrument is difficult to create ... because the universe of possible content is virtually infinite" (page 150). Again, referring to Cronbach [53], Straub recommends an expert to evaluate the instruments. This recommendation for establishing content validity for survey instruments with the help of expert assessment has been adopted in the recommendation of focus groups [54] for instantiation validity by Lukyanenko et al. [37] .

We extend this suggestion of sampling object stimuli (experimental or questionnaire items) to the sampling of artifacts and features in ITDR. As mentioned earlier, the problems of IV, while present in other disciplines, are particularly important for studies involving IT. Unlike simple drawings, silhouettes, stick figures, etc. common in psychology e.g., [55], [56], IT are more complex. The patterns of interaction with IT are also constantly evolving, further confounding efforts to detect extraneous interferences.

3. Artifact Sampling Method

Motivated by the methodological suggestions and arguments in social sciences, here we develop a preliminary method for ***artifact sampling***. The artifact sampling method should be used during the evaluation phase of ITDR and mainly focuses on the selection of the artifacts for the study.

Artifact sampling extends the concept of stimulus sampling from experimental psychology and scale reliability from survey research to research involving software artifacts. Artifact sampling entails selecting multiple artifacts from the space of valid design possibilities. Software artifacts are intended to instantiate, through certain features, a particular level of one or more theoretical constructs, for example a high degree of personalization. Given the typically very large design space of design features, sampling from this design space produces a set of artifacts representative of the desired theoretical construct level, e.g. high personalization.

Instrument validation in survey research establishes construct validity by answering the question whether "instruments show stability across methodologies". In other words, construct validity "asks whether the measures chosen are ... merely artifacts of the [measurement] methodology itself" [52, p. 150]. The immediate parallel in instantiation validity is the question whether the instantiation is biased by its construction methodology [37]. To answer this question, different artifacts may be sampled from different construction methods (e.g. web-based, mobile app), interface paradigms (e.g. mouse, touch, VR), or application domains (e.g. financial services, social networking, e-commerce) to enable identifying the influence of any of these factors on the artifact as necessary to ensure the external validity claimed by the researchers.

These ideas form the basis for the proposed method, the steps of which are outlined below. Once the theoretical sample space is established, sampling procedures should be applied to select multiple artifacts, which can then be implemented and used for evaluation. Next, we propose steps to be followed in artifact sampling.

Step 1: From theory to instantiation space.

The success of artifact sampling begins with the theoretical rigor in a study. We recommend to clearly and precisely define the theoretical construct that corresponds to the features of the IT. A clear definition is necessary for the construction of an appropriate IT artifact. Based on the theoretical definition, create the theoretical ***instantiation space*** by identifying necessary and sufficient features and deriving from these a conceptual space of valid implementations.

To illustrate, consider again the theoretical context of IT adoption, and a researcher hypothesizing that IT with "high social interactivity" (a theoretical construct) results in higher adoption by users. This research would start with a clear and precise definition of the focal construct of interest, considering the existing body of knowledge that pertains to the construct and ways it has been operationalized in the past [57], [58].

The specific theoretical features should then be used to derive a multitude of possible designs corresponding to specific ways this construct may be implemented in line with the proposed construct definition. This first entails constructing an ***instantiation space*** by closely examining the theory and deriving from it a conceptual space of valid implementations. The process of identifying a theoretical space and deriving multiple objects that instantiate it is becoming better understood in psychology, as it develops stimuli libraries (e.g., [59], [60]). From this work, it is evident that the process requires theoretical rigor, as it involves developing a thorough understanding of what makes an implementation a valid instance of the construct [50]. Here, design science research in IS, in particular, stands to inform artifact sampling, as it has a tradition of working with artifacts at instantiated and conceptual levels [35], [37], [59]–[63].

These implementations need not consider every possible way to implement the construct (now and in the future) but, as argued by Wells and Windschitl [51, p. 1115], should be representative enough and contain enough variation to capture as many possible confounds as feasible for the project [38], see, [66]. Constructing an instantiation space therefore requires both deep understanding of the construct and of the design possibilities [36].

Returning to the "high social interactivity" construct example, researchers might conceive various ways to implement this construct in a website using different construction methods, interface elements, and application domains. The instantiation space in this

example is a conceptual space that can be represented as a matrix of specific features (e.g., red font color for H1 heading), of feature dimensions (e.g., font size, background color, navigation structure) that correspond to each artifact deemed to be a valid instantiation of the focal theoretical construct (see Table 1).

Table 1. Instantiation Space Matrix

IT Artifact	*Dimension 1*	*Dimension 2*	*Dimension N*
A_1	Feature 1	Feature 3	Feature 6
A_2	Feature 1	Feature 4	Feature 6
A_3	Feature 1	Feature 4	Feature 7
A_N	Feature 2	Feature 5	Feature 7

The feature dimensions are derived from the focal theoretical construct by determining which design features are necessary and sufficient to convey through design the essence of the construct.

The specific features are chosen in two ways:

1. When no or very few instances of the focal theoretical construct exist, it should be based on how a given feature dimension can be potentially be realized in a real-world IT. For example, the font color dimension for a web-based IT can be realized through any of the web-safe colors in a color palette.
2. When there are existing IT artifacts, by examining real-world instances of the focal theoretical construct (i.e., existing applications that are available and deemed by researchers to be examples of the theoretical construct of interest). For example, there could be existing websites that exhibit high degree of social interactivity (e.g., Facebook.com, Instagram.com, Twitter.com). The researchers then examine each of the real-world projects to extract specific features for the feature dimensions identified based on the focal theoretical construct (e.g., Table 2).

Table 2. Sample Instantiation Space Matrix

IT Artifact	*Rapid notifications*	*Has Live Chat*	*Network nature*
Facebook.com	Yes	Yes	Friends-focused
Instagram.com	Yes	No	Photo/video-focused
Twitter.com	Yes	No	Information-focused

Step 2: Determine the nature of the sample.

Once the instantiation space is established, use it to select (when there are accessible existing IT applications) or create specific permutations of the artifacts. Since it may be impractical to create or use every valid IT instance, we suggest sampling from the instantiation space.

Sampling from the instantiation space may be pursued in two principal ways:

1. Sampling for artifact diversity; and/or
2. Sampling for artifact homogeneity.

First, researchers can sample for artifact diversity and breadth to cover many points in the design space. The aim here is to improve generalizability (i.e. inference to the population) and get an assessment of the heterogeneity of the design space (which will inform any generalizability claims one makes). Researchers can use the instantiation space matrix (e.g., Table 1), and select artifacts that have different features along the feature dimensions, such that every unique feature is represented in the sample.

In the second case, researchers sample very similar points in the design space for homogeneity to get a more reliable sample and reliable theoretical claims. Here, the aim is homogeneity of the sample so that minor local variations of the design space "cancel each other out". For example, researchers may consider artifacts with the most similar features for each feature dimension in the instantiation space matrix.

Finally, researchers may combine the two strategies above to obtain heterogenous set of homogenous sets of samples, which would allow reliable claims about each sample point and also allow claims to generalize based on a thorough understanding of the different parts of the design space. Thus, we recommend combined approaches to the extent possible.

Step 3: Sampling.

Implement a sampling procedure by drawing from the instantiation space. The sample size and its selection is naturally constrained by:

- (expected) natural variability of relevant features in the population of artifacts (where greater variability calls for more artifacts);
- expected confounding factors and the difficulty in detecting and controlling (here, more artifacts could be used, at least, partially to assuage concerns about potential confounds);
- desire to draw stronger inferences (which may suggest striving for larger sample sizes and

random selection to perform analysis of variance tests over groups of artifacts); and

- pragmatic considerations (e.g., cost and effort of implementation, which may limit the number of artifacts).

One suggestion is to echo the sentiments from multi-item measurement for survey scale development that encourages 3-7 items per construct to achieve adequate reliability [67]. While this guidance is tentative at best, it provides a starting point to compare to a single instantiation.

Step 4: Evaluate each artifact-condition.

The objective of artifact sampling is to convert each artifact into an object of evaluation. When possible, each artifact becomes a separate experimental condition. This means that for every artifact-condition, researchers would need to provide an appropriate evaluation procedure.

For example, for each artifact that corresponds to the "high social interactivity", researchers may choose to utilize an experimental design. This means that a large pool of participants would be randomly assigned to each artifact-condition (e.g., 20 per artifact), and the participants would be asked to experience the artifact and then respond to a set of questions (e.g., asking about the intentions to use the system) which would be common across all the artifacts (conditions).

Clearly, assigning a separate group of participants for each artifact would require a large pool of participants, and may not be realistic for all projects. Pragmatic considerations, such as availability of research participants, may result in a different study design (e.g., asking participants to experience multiple artifacts per session). The choice of strategy here ultimately depends on the available resources and the intention to draw stronger inferences from the results.

Step 5: Analyze results and draw conclusions.

Analyze the results by condition and in aggregate. Here, the presence of multiple artifacts (and the corresponding multiple experimental conditions) can be used in a variety of ways. For example, researchers can report on the general convergence or divergence between different experimental conditions.

To illustrate, consider two possibilities shown in Figure 1. In Scenario 1, we see consistent results across the different conditions in which different variations on the same construct were used (here, the artifact intends to instantiate a theoretical construct of "high social interactivity"). From the results obtained, a researcher can be quite confident in the overall conclusion that employing IT that exhibits high social interactivity results in increased adoption of the underlying technology by users. In contrast, if the results instead are more similar to Scenario 2, such conclusion, if drawn, should be qualified. Furthermore, the inconsistent behavior between different IT systems (all purporting to instantiate the same underlying construct), may suggest that there could be unforeseen and potentially unknown confounding factors - features of the technology. If possible, a deep probing into the design features of the artifacts that do not behave in the expected manner would be advisable, as this could potentially produce new knowledge and enrich our understanding of the underlying theoretical construct of interest.

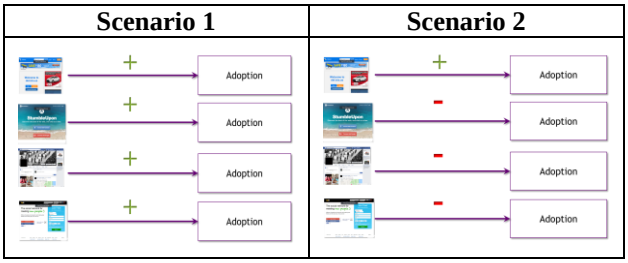

Figure 1. Two alternative scenarios of a hypothetical artifact sampling

When possible, researchers may conduct additional statistical analysis on the extent of convergence or divergence between the conditions corresponding to each sampled artifact. For example, researchers may use Cronbach alpha as a numerical index of concordance.

3. Future Work

In this paper, we propose a novel methodological concept – ***artifact sampling*** – intended to increase both inner and outer instantiation validity of ITDR studies involving software artifacts. It helps to address inner or operationalization validity by helping to mitigate potential confounds due to the complexity of IT artifacts. It also aids in establishing outer or conclusion validity by offering richer empirical evidence to draw upon and providing for stronger inferences and conclusions.

Artifact sampling is an important methodological practice that stands to increase rigor in research dealing with software artifacts. Nevertheless, we suggest it is not being adequately undertaken in ITDR research to date, potentially biasing conclusions of studies that rely on artifacts.

The key contribution of this research is to motivate future work on the method of artifact sampling. We pave the way for future work by providing the foundation for artifact sampling. In particular, artifact

sampling has foundations in sampling theory, the notion of representative design, and is akin to the well-established norms for increasing reliability in psychometric research. The idea of artifact sampling is becoming increasingly accepted in psychology, where it is known as stimuli sampling. Recently, renewed and stronger arguments in favor of stimuli sampling have been made and new approaches and methods are being proposed. Libraries of stimuli are also proliferating at a rapid rate (see references above), thus further underscoring the on-going acceptance of the idea. This motivated us to consider the implications of these developments for ITDR, culminating in our artifact sampling proposal.

Admittedly, the artifact sampling method should be assessed and revised. We acknowledge such limitations in our current proposal and call for more research to help provide specific guidelines. First, artifact sampling may not be always be useful, just as in some cases a single-item survey scale is sufficient [67]. For example, artifact sampling may not be needed if testing the effect of Facebook use (as a social network site) if there is no intent to generalize to other social media technologies. Indeed, sampling potential social network artifacts may not be practical or useful in such situations. Likewise, if an artifact has wide acceptance, it may be useful to study its effects without sampling. Artifact sampling is more geared toward nomothetic rather than idiographic research objectives [30], [68]–[70]. Second, guidelines on how to establish the instantiation space are needed to help researchers carefully plan out their instantiation options. The dimensions of the design space should be orthogonal, as much as possible, to ensure that the sampled artifacts are independent. Third, guidelines are needed for establishing the independence of the sampled items as well as the number of items necessary. Fourth, the development of quantitative or qualitative techniques that allow subjects to evaluate the instantiation validity of objects is necessary.

Clearly, artifact sampling will not apply to cases where the instantiation space is limited and small and where the dimensions of the space cannot be defined independently of each other. However, as argued in [19], [38], many ITDR research questions deal with situations where it is unclear how to design an artifact and many (and sometimes potentially an unlimited number of) design choices exist. Indeed, the notion of a potentially vast space of possible operationalizations is recognized in other disciplines [51], and we believe it should at least be considered in ITDR, especially during the process of designing and evaluating artifacts. Importantly, however, this process elevates IT-based research to higher levels of rigor as it helps to addresses instantiation validity concerns and increase the confidence in the conclusions of ITDR studies. It also opens a variety of novel and intriguing methodological possibilities, promising better science and advancing IT design knowledge.

We acknowledge that the notion of artifact sampling for instantiation validity might be met with its own criticisms. For example, some may argue that design decisions are ultimately guided by theory, and not empirical evaluation (a position we also hold, but we suggest that often it is difficult to settle on a single correct design). Drawing from the methodological context of scale development research, the choice of whether to drop/add an item is ultimately determined by theoretical reasons, not just the empirical evaluation. However, empirical measurement model techniques do provide recommendations with respect to how valid the measurement of the construct is with the presence/absence of the item. Another criticism may be the notion of a program of study [71] and/or replication of a design to ultimately find the appropriate operationalization [72]. For example, perhaps in the initial operationalization of the design, providing a definitive theoretical justification for design choices is impractical, and, further studies can help refine and confirm the validity of the design choices e.g., [54]. We believe this approach is also sound, but note that much of ITDR research has been criticized for the lack of extensive replication and some question whether a cumulative research tradition is even possible when dealing with ever changing IT artifacts [19]. Future studies should explore in greater detail when artifact sampling is more effective and epistemically appropriate, and when other strategies should be pursued.

In the future, we hope to better understand the process of artifact sampling, develop best practices, address the issue of when this method should be applied and provide specific examples that illustrate application of this concept. Once the notion of instantiation validity is well defined, and all aspects of the artifact sampling method are established, future research should conduct empirical evaluations to demonstrate empirically the concerns related to instantiation validity as well as evaluations of the proposed artifact sampling method as a solution to these issues.

We also hope that this paper will motivate further discussions about both the proposed idea of artifact sampling and the broader concerns of instantiation validity.

5. References

[1] L. Aksoy, P. N. Bloom, N. H. Lurie, and B. Cooil, “Should Recommendation Agents Think Like

People?," *J. Serv. Res.*, vol. 8, no. 4, pp. 297–315, May 2006.
[2] I. Benbasat and W. Wang, "Trust In and Adoption of Online Recommendation Agents," *J. Assoc. Inf. Syst.*, vol. 6, no. 3, pp. 72–101, 2005.
[3] S. X. Komiak and I. Benbasat, "Understanding Customer Trust in Agent-Mediated Electronic Commerce, Web-Mediated Electronic Commerce, and Traditional Commerce," *Inf. Technol. Manag.*, vol. 5, no. 1, pp. 181–207, 2004.
[4] M. Kosmala, A. Wiggins, A. Swanson, and B. Simmons, "Assessing data quality in citizen science," *Front. Ecol. Environ.*, vol. 14, no. 10, pp. 551–560, 2016.
[5] A. Wiggins and K. Crowston, "From Conservation to Crowdsourcing: A Typology of Citizen Science," in *44th Hawaii International Conference on System Sciences*, 2011, pp. 1–10.
[6] J. Prpić, P. P. Shukla, J. H. Kietzmann, and I. P. McCarthy, "How to work a crowd: Developing crowd capital through crowdsourcing," *Bus. Horiz.*, vol. 58, no. 1, pp. 77–85, 2015.
[7] R. Lukyanenko, J. Parsons, and Y. Wiersma, "Emerging problems of data quality in citizen science," *Conserv. Biol.*, vol. 30, no. 3, pp. 447–449, 2016.
[8] C. Cardamone *et al.*, "Galaxy Zoo Green Peas: discovery of a class of compact extremely star-forming galaxies," *Mon. Not. R. Astron. Soc.*, vol. 399, no. 3, pp. 1191–1205, 2009.
[9] C. T. Callaghan and D. E. Gawlik, "Efficacy of eBird data as an aid in conservation planning and monitoring," *J. Field Ornithol.*, vol. 1, no. 1, pp. 1–7, 2015.
[10] R. Lukyanenko, J. Parsons, Y. F. Wiersma, G. Wachinger, B. Huber, and R. Meldt, "Representing Crowd Knowledge: Guidelines for Conceptual Modeling of User-generated Content," *J. Assoc. Inf. Syst.*, vol. 18, no. 4, pp. 297–339, 2017.
[11] R. Lukyanenko, J. Parsons, and Y. Wiersma, "The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-generated Content," *Inf. Syst. Res.*, vol. 25, no. 4, pp. 669–689, 2014.
[12] I. Benbasat, *Laboratory experiments in information systems studies with a focus on individuals: a critical appraisal*, vol. 2. Cambridge, MA: Harvard Business School, 1989.
[13] M.-C. Boudreau, D. Gefen, and D. W. Straub, "Validation in Information Systems Research: A State-of-the-Art Assessment," *MIS Q.*, vol. 25, no. 1, pp. 1–16, 2001.
[14] A. Burton-Jones, Y. Wand, and R. Weber, "Guidelines for Empirical Evaluations of Conceptual Modeling Grammars," *J. Assoc. Inf. Syst.*, vol. 10, no. 6, pp. 495–532, 2009.
[15] W. J. Orlikowski and C. S. Iacono, "Research commentary: Desperately seeking the IT in IT research—a call to theorizing the IT artifact," *Inf. Syst. Res.*, vol. 12, no. 2, pp. 121–134, 2001.
[16] J. Parsons and L. Cole, "What do the pictures mean? Guidelines for experimental evaluation of representation fidelity in diagrammatical conceptual modeling techniques," *Data Knowl. Eng.*, vol. 55, no. 3, pp. 327–342, 2005.
[17] D. Straub, M.-C. Boudreau, and D. Gefen, "Validation guidelines for IS positivist research," *Commun. Assoc. Inf. Syst.*, vol. 13, no. 1, pp. 1–63, 2004.
[18] T. D. Cook and D. T. Campbell, *Quasi-experimentation : design & analysis issues for field settings*. Chicago: Rand McNally College Pub. Co., 1979.
[19] R. Lukyanenko, J. Evermann, and J. Parsons, "Instantiation Validity in IS Design Research," in *DESRIST 2014, LNCS 8463*, Springer, 2014, pp. 321–328.
[20] J. Iivari, "A Paradigmatic Analysis of Information Systems as a Design Science," *Scand. J. Inf. Syst.*, pp. 39–64, 2007.
[21] C. Avgerou, "The significance of context in information systems and organizational change," *Inf. Syst. J.*, vol. 11, no. 1, pp. 43–63, 2001.
[22] E. Davidson and M. Chiasson, "Contextual influences on technology use mediation: a comparative analysis of electronic medical record systems," *Eur. J. Inf. Syst.*, vol. 14, no. 1, pp. 6–18, 2005.
[23] G. Johns, "The essential impact of context of organizational behavior," *Acad. Manage. Rev.*, vol. 31, no. 2, 2006.
[24] A. Burton-Jones and O. Volkoff, "How can we develop contextualized theories of effective use? A demonstration in the context of community-care electronic health records," *Inf. Syst. Res.*, vol. Forthcoming, pp. 1–40, 2017.
[25] J. Evermann, "Contextual factors in database integration—a delphi study," presented at the International Conference on Conceptual Modeling, 2010, pp. 274–287.
[26] W. Hong, F. K. Chan, J. Y. Thong, L. C. Chasalow, and G. Dhillon, "A framework and guidelines for context-specific theorizing in information systems research," *Inf. Syst. Res.*, vol. 25, no. 1, pp. 111–136, 2013.
[27] A. Hevner and S. Chatterjee, *Design Research in Information Systems: Theory and Practice*, vol. 22. Springer, 2010.
[28] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Q.*, vol. 37, no. 2, pp. 337–355, 2013.

[29] A. Hevner, S. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004.
[30] R. L. Baskerville, M. Kaul, and V. C. Storey, "Genres of Inquiry in Design-Science Research: Justification and Evaluation of Knowledge Production.," *MIS Q.*, vol. 39, no. 3, pp. 541–564, 2015.
[31] S. T. March and G. F. Smith, "Design and natural science research on information technology," *Decis. Support Syst.*, vol. 15, no. 4, pp. 251–266, Dec. 1995.
[32] J. F. Nunamaker, M. Chen, and T. D. Purdin, "Systems development in information systems research," *J. Manag. Inf. Syst.*, vol. 7, no. 3, pp. 89–106, 1991.
[33] J. F. Nunamaker Jr and R. O. Briggs, "Toward a broader vision for information systems," *ACM Trans. Manag. Inf. Syst. TMIS*, vol. 2, no. 4, p. 20, 2011.
[34] J. Akoka, I. Comyn-Wattiau, N. Prat, and V. C. Storey, "Evaluating Knowledge Types in Design Science Research: An Integrated Framework," presented at the International Conference on Design Science Research in Information Systems, 2017, pp. 201–217.
[35] J. Venable, J. Pries-Heje, and R. Baskerville, "A comprehensive framework for evaluation in design science research," in *DESRIST 2012, LNCS 7286*, Springer, 2012, pp. 423–438.
[36] O. Arazy, N. Kumar, and B. Shapira, "A theory-driven design framework for social recommender systems," *J. Assoc. Inf. Syst.*, vol. 11, no. 9, pp. 455–490, 2010.
[37] R. Lukyanenko, J. Evermann, and J. Parsons, "Guidelines for Establishing Instantiation Validity in IT Artifacts: A Survey of IS Research," in *DESRIST 2015, LNCS 9073*, Berlin / Heidelberg: Springer, 2015.
[38] R. Lukyanenko and J. Parsons, "Reconciling theories with design choices in design science research," in *DESRIST 2013, LNCS 7939*, Springer Berlin / Heidelberg, 2013, pp. 165–180.
[39] L. Chandra, S. Seidel, and S. Gregor, "Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions," *Hawaii Int. Conf. Syst. Sci.*, pp. 4039–4047, 2015.
[40] S. L. Lohr, *Sampling: Design and Analysis*. Cengage Learning, 2009.
[41] K. R. Hammond and T. R. Stewart, *The essential Brunswik: Beginnings, explications, applications.* Oxford University Press, 2001.
[42] E. Brunswik, "Representative design and probabilistic theory in a functional psychology.," *Psychol. Rev.*, vol. 62, no. 3, p. 193, 1955.
[43] E. Brunswik, "Organismic achievement and environmental probability.," *Psychol. Rev.*, vol. 50, no. 3, p. 255, 1943.
[44] S. Highhouse, "Designing experiments that generalize," *Organ. Res. Methods*, vol. 12, no. 3, pp. 554–566, 2009.
[45] G. Lindzey, D. Gilbert, and S. T. Fiske, *The handbook of social psychology*. Oxford University Press, 1998.
[46] D. R. Bonge, W. J. Schuldt, and Y. Y. Harper, "The experimenter-as-fixed-effect fallacy," *J. Psychol.*, vol. 126, no. 5, pp. 477–486, 1992.
[47] M. K. Dhami, R. Hertwig, and U. Hoffrage, "The role of representative design in an ecological approach to cognition.," *Psychol. Bull.*, vol. 130, no. 6, p. 959, 2004.
[48] G. A. Fontenelle, A. P. Phillips, and D. M. Lane, "Generalizing across stimuli as well as subjects: A neglected aspect of external validity.," *J. Appl. Psychol.*, vol. 70, no. 1, p. 101, 1985.
[49] L. Kelley, *Issues, theory, and research in industrial/organizational psychology*, vol. 82. Elsevier, 1992.
[50] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity," *J. Exp. Psychol. [Hum. Learn.]*, vol. 6, no. 2, pp. 174–215, 1980.
[51] G. L. Wells and P. D. Windschitl, "Stimulus sampling and social psychological experimentation," *Pers. Soc. Psychol. Bull.*, vol. 25, no. 9, pp. 1115–1125, 1999.
[52] D. W. Straub, "Validating Instruments in MIS Research," *MIS Q.*, vol. 13, no. 2, pp. 147–169, 1989.
[53] L. J. Cronbach, "Test validation," in *Educational measurement*, R. Thorndike, Ed. 1971, pp. 443–507.
[54] M. C. Tremblay, A. R. Hevner, and D. J. Berndt, "Focus Groups for Artifact Refinement and Evaluation in Design Research.," *Commun. Assoc. Inf. Syst.*, vol. 26, pp. 599–618, 2010.
[55] H. O. de Beeck and J. Wagemans, "Visual object categorisation at distinct levels of abstraction: A new stimulus set," *Perception*, vol. 30, no. 11, pp. 1337–1361, 2001.
[56] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyesbraem, "Basic Objects in Natural Categories," *Cognit. Psychol.*, vol. 8, no. 3, pp. 382–439, 1976.
[57] K. Larsen and C. H. Bong, "A Tool for Addressing Construct Identity in Literature Reviews and Meta-Analyses," *MIS Q.*, vol. 40, no. 3, pp. 1–23, 2016.
[58] J. Endicott, K. R. Larsen, R. Lukyanenko, and C. H. Bong, "Integrating Scientific Research: Theory and Design of Discovering Similar Constructs," in *AIS*

SIGSAND Symposium, Cincinnati, Ohio, 2017, pp. 1–7.
[59] F.-X. Alario and L. Ferrand, "A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition," *Behav. Res. Methods Instrum. Comput.*, vol. 31, no. 3, pp. 531–552, 1999.
[60] S. Berman, D. Friedman, M. Hamberger, and J. G. Snodgrass, "Developmental picture norms: Relationships between name agreement, familiarity, and visual complexity for child and adult ratings of two sets of line drawings," *Behav. Res. Methods Instrum. Comput.*, vol. 21, no. 3, pp. 371–382, 1989.
[61] O. Arazy and R. Kopak, "On the measurability of information quality," *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 1, pp. 89–99, 2011.
[62] L. Chandra Kruse, S. Seidel, and S. Purao, "Making Use of Design Principles," presented at the DESRIST 2016, LNCS 9661, 2016, pp. 37–51.
[63] S. Gregor, O. Müller, and S. Seidel, "Reflection, Abstraction And Theorizing In Design And Development Research.," presented at the ECIS, 2013, p. 74.
[64] N. Prat, I. Comyn-Wattiau, and J. Akoka, "A Taxonomy of Evaluation Methods for Information Systems Artifacts," *J. Manag. Inf. Syst.*, vol. 32, no. 3, pp. 229–267, 2015.
[65] S. Purao, "Truth or dare: The ontology question in design science research," *J. Database Manag. JDM*, vol. 24, no. 3, pp. 51–66, 2013.
[66] R. Baskerville and J. Pries-Heje, "Design Theory Projectability," in *Information Systems and Global Assemblages.(Re) Configuring Actors, Artefacts, Organizations*, Springer, 2014, pp. 219–232.
[67] A. Diamantopoulos, M. Sarstedt, C. Fuchs, P. Wilczynski, and S. Kaiser, "Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective," *J. Acad. Mark. Sci.*, vol. 40, no. 3, pp. 434–449, 2012.
[68] R. Lukyanenko and D. Darcy, "On Systematicity: Expanding the Diversity of Design Science Research Contributions," in *DESRIST 2016, Tackling Society's Grand Challenges with Design Science*, St. John's, NL Canada, 2016, pp. 1–7.
[69] A. Amrollahi, R. Lukyanenko, and A. Castellanos, "Multi-Paradigmatic Theorizing: Mixing Design and Exploration.," in *AMCIS*, 2017.
[70] K. R. Larsen and D. S. Hovorka, "Developing interfield nomological nets," in *System Science (HICSS), 2012 45th Hawaii International Conference on*, 2012, pp. 5194–5203.
[71] I. Lakatos and A. Musgrave, *Criticism and the Growth of Knowledge*. Cambridge England ; New York: Cambridge University Press., 1970.
[72] A. R. Dennis and J. S. Valacich, "A replication manifesto," *AIS Trans. Replication Res.*, vol. 1, no. 1, p. 1, 2014.