Interactions Between System Evaluation and Theory Testing: A Demonstration of the Power of a Multifaceted Approach to Information Systems Research

JINWEI CAO, JANNA M. CREWS, MING LIN, AMIT DEOKAR, JUDEE K. BURGOON, AND JAY F. NUNAMAKER JR.

JINWEI CAO is an Assistant Professor in the Department of Accounting and Management Information Systems at the University of Delaware. She received her Ph.D. in Management Information Systems from the University of Arizona. Prior to joining the University of Delaware, she was a Research Associate at the Center for the Management of Information and the MIS department at the University of Arizona (2000–2005). Her research interests include technology supported learning, multimedia information systems, human–computer interaction and usability, and knowledge management.

JANNA M. CREWS is an Assistant Professor in the Department of Accounting and Information Systems at the University of Nevada, Reno. She received her Ph.D. from the University of Arizona in 2004, where she was a Research Associate for the Center for the Management of Information (CMI). Her primary research interests include the design of technology-mediated learning (TML), computer-assisted instruction (CAI), distance learning, multimedia training, IT education, human–computer interaction, usability, and utilization of collaboration technologies. A Certified Public Accountant (CPA), Dr. Crews holds Bachelor and Master of Accountancy degrees from the University of Oklahoma and spent several years working in public accounting, banking and real estate industries.

MING LIN is a Ph.D. Candidate in the Management Information Systems Department at the University of Arizona. He received his M.S. in Computer Science from the Institute of Computing Technology, Chinese Academy of Sciences, and his B.S. in Computer Science from Northern Jiaotong University, China. His current research interests include knowledge management, computer-assisted learning, and decision support systems. His current research focuses on developing multimedia systems and studying their effects on learning and knowledge management. He has also been an active researcher in several projects funded by NSF, Ford Foundation, and DARPA.

AMIT DEOKAR is a Ph.D. Candidate in the Management Information Systems Department at the University of Arizona. He earned an M.S. in Industrial Engineering from the University of Arizona and a B.S. in Mechanical Engineering from the V.J. Technological Institute, University of Mumbai, India. His research interests include col-
laboration systems, business process management systems, and artificial intelligence. He is also a Research Associate for the Center for the Management of Information (CMI) and the MIS Department at the University of Arizona, where he has done research on computer-aided deception detection tools, content management systems, and knowledge management systems.

JUDEE K. BURGOON is a Professor of Communication, Family Studies, and Human Development, a Director of Human Communication Research for the Center for the Management of Information, and an Associate Director of the Media Interface Network Design Laboratory at the University of Arizona. Among her research interests are interpersonal and nonverbal communication, communication theory, and research methods. Her current research centering on detection of deception and intent through computer-assisted tools has been funded by the Departments of Defense and Homeland Security.

JAY F. NUNAMAKER JR. is Regents and Soldwedel Professor of MIS, Computer Science, and Communication and Director of the Center for the Management of Information at the University of Arizona, Tucson. He received his Ph.D. in Systems Engineering and Operations Research from Case Institute of Technology, an M.S. and B.S. in Engineering from the University of Pittsburgh, and a B.S. from Carnegie Mellon University. Dr. Nunamaker received the LEO Award from the Association of Information Systems at the International Conference on Information Systems (ICIS) in Barcelona, Spain, December 2002. This award is given for a lifetime of exceptional achievement in information systems. He was elected as a fellow of the Association of Information Systems in 2000. He has over 40 years of experience in examining, analyzing, designing, testing, evaluating, and developing information systems. He has served as a test engineer at the Shippingport Atomic Power facility, as a member of the ISDOS team at the University of Michigan, and as a member of the faculty at Purdue University, prior to joining the faculty at the University of Arizona in 1974. His research on group support systems addresses behavioral as well as engineering issues and focuses on theory as well as implementation. He has been a licensed professional engineer since 1965.

ABSTRACT: Historically, information systems (IS) researchers have questioned which research paradigms, activities, and methods IS research should follow. In this paper, we argue that different research methods and activities may interact with each other, different research paradigms may complement each other due to such interactions, and therefore, a multimethodological, cross-paradigm research approach may result in better IS research than a singular approach. Three existing multimethodological IS research frameworks are reviewed and summarized into an integrated approach. Two types of interactions between different research methods across system evaluation and theory testing research activities are identified. A three-year research study about a computer-based training system for deception detection (Agent99 Trainer) provides a concrete example to demonstrate the existence and research benefits of these two types of interactions, as well as the benefits of a multimethodological, cross-paradigm IS research approach.

KEY WORDS AND PHRASES: IS research framework, research activity, research method, research paradigm.
RESEARCH ON INFORMATION SYSTEMS (IS) has been evolving since the 1960s. “The mission of Information Systems research is to study the effective design, delivery, use and impact of information technologies [IT] in organizations and society” [19, p. 3]. The inclusion of information, technology, and organization in this articulation serves to illustrate the complex, interdisciplinary nature of IS research. Because of its roots in multiple disciplines, such as computer science, management, and communication, the IS research community continues to question what the core features of IS research are and how IS research should be conducted. Many of these discussions concentrate on which research paradigms, activities, and methods are appropriate for conducting IS research [17]. In this paper, we posit that different research methods and activities may interact with each other in certain beneficial ways, different research paradigms may complement each other because of such interactions, and therefore, a multimethodological, cross-paradigm research approach may result in better IS research.

Recently, Hevner et al. [15] focused the discussion in IS research on two major research paradigms—natural science and design science. Although some equate research paradigms with research epistemology (i.e., positivist, interpretive, and critical), in this paper, we follow the definition in Hevner et al. [15], which refers to the two research paradigms of natural (or behavioral) science and design science. Under the behavioral science paradigm, researchers aim to understand phenomena related to the development and use of IS, whereas under the design science paradigm, researchers aim to develop IS artifacts and improve their performance. Paradigms in this context thus concern the substance of “what is being studied” rather than the epistemology, or “way of knowing,” used to understand the phenomena of interest [16]. For example, a positivist research approach seeks to discover and validate propositional statements about a given phenomenon, an interpretive research approach seeks to understand the meanings assigned by people who interact with the phenomenon, and a critical research approach seeks to evaluate and transform the dominant understanding of the phenomenon [34]. Research under both the behavioral science paradigm and the design science paradigm can follow one or more of these underlying epistemologies. In this paper, our focus is on the differences and interactions across the behavioral science and design science research paradigms, but is not limited to any specific epistemologies.

March and Smith summarized four basic activities (theorize, justify, build, and evaluate) for conducting IS research under these two research paradigms (behavioral science and design science) [23]. Behavioral science consists of the first two activities, theorize and justify/test (theories), and design science consists of the other two, build and evaluate (artifacts). Each of these four research activities may involve many specific, detailed research methods that may cross paradigms. For example, a controlled experiment is a research method commonly applied in theory justification activities of the behavioral science paradigm, as well as in artifact evaluation activities of the design science paradigm. Note that, although some may distinguish method
from methodology [27], we use these two terms interchangeably to refer to “well-defined sequences of operations that if carried out proficiently yield predictable results” [27, p. 241].

For many years, a division has existed among IS researchers [15, 23] based on these two research paradigms. Furthermore, many, if not most, IS research studies were conducted with single or very limited research methods under one paradigm. In fact, research shows that more than 80 percent of individual IS research studies in the 1990s only used a single research method [28, 30], such as case study, experiment, or survey [1]. This single-method approach, although usually easier to implement, may result in a narrow, inconclusive explanation of a particular research problem [10].

After years of discussions about which paradigm and which methods constitute legitimate scientific research in IS, there is an emerging recognition of the dual nature of IS research and an effort toward transforming the bifurcation of methods into a grounded, complementary research foundation [14, 15, 23]. We agree with March and Smith [23] that, although we should still accept the differences between the design science and behavioral science research paradigms and their inherent research activities, we should recognize that the interactions between these research activities are important and the resulting complementary research cycle between behavior science and design science is the key for IS research success.

Starting in the 1990s, a number of IS research frameworks have been proposed to guide the combination of multiple research methods, activities, and paradigms in IS research [15, 23, 27, 32, 33]. These frameworks suggest that triangulation of different research methods or activities has the potential to lead to more powerful and insightful findings. This paper will further argue that the different research methods relate to and inform each other, and that the benefits deriving from the interactions among these methods outweigh the costs in time and effort of implementing such a multimethodological approach.

In proposing such interactions among research methods and promoting for an integrated IS research framework, we draw upon our experience with a recent research study (the Agent99 [A99] study described in the fourth section), which, in turn, was grounded in Nunamaker et al.’s multimethodological IS research framework [32, 33]. Using the A99 study as a concrete example, we explain two types of interactions that may occur between the research methods used to “evaluate” (system) research activities (under the design science paradigm) and “justify” (theory) research activities (under the behavioral science paradigm). In one interaction, the interpretation of results from one method may affect the interpretation of results from the other, and vice versa. In the other interaction, the interpretation of results from one method may affect the design of another method. We demonstrate via the A99 study example that, through these interactions, the design science and behavioral science paradigms complement each other, increasing the validity of IS research and improving the resulting theories and artifacts.
Multimethodological IS Research—The Frameworks and Research Activities

As noted, more and more researchers have begun to recognize the importance of combining multiple research methods, activities, or paradigms in IS research. From the IS research literature, we found three frameworks proposing guidelines for multimethodological, cross-paradigm IS research.

The Existing Research Frameworks

In 1991, Nunamaker et al. [33] advocated a multimethodological approach to IS research on system development. Their research framework includes four complementary research strategies—theory building, systems development, observation, and experimentation (Figure 1). Each research strategy may be investigated by a number of complementary research methods that provide valuable feedback to one another, within and across the different strategies. For example, the observation strategy may be investigated by methods including, but not limited to, case study, survey, and field study. The results from an observation strategy may help researchers to formulate specific hypotheses to be tested by an experimentation method. In the Nunamaker et al. framework, system development is the hub of IS research and interacts with other research strategies to form an integrated research program.

The four research strategies in the Nunamaker et al. [33] framework can be mapped onto the four research activities proposed in March and Smith’s [23] later framework for IS research (Figure 2). This framework emphasizes that IS research should be concerned both with utility, from a design science perspective, and with theory, from a behavioral science perspective. Successful conduct of IS research needs both design science research activities (build and evaluate) and behavioral science research activities (theorize and justify). The basic questions that each activity tries to answer can be summarized as follows. Building an artifact demonstrates that such an artifact can be constructed and answers the basic question, “Does it work?” Evaluating the artifact tries to determine the performance of the artifact by answering the question, “How well does it work?” Given an artifact whose performance has been evaluated, the activity of theorizing brings up the question, “What natural laws can explain how or why the artifact works within its environment?” Finally, justifying theories answers the question, “Can these natural laws explain how or why the artifact works within its environment?”

This framework is similar to the Nunamaker et al. framework in that the “build” activity can be mapped to the system development strategy in the Nunamaker et al. framework. In addition, the “theorize” activity maps to the theory building strategy, while the “evaluate” and “justify” activities map to the observation and experimentation strategies, respectively. March and Smith also depict the interactions between the research activities of building an artifact (i.e., the system) and theorizing about that artifact. That is, building an IT artifact creates an object for or provides proof of
Figure 1. Nunamaker et al.'s Multimethodological Approach to IS Research [33]

Figure 2. March and Smith's Research Framework in IT [23]. Reprinted with permission from Elsevier.
the claims generated from the theorizing activity, while the theories proposed by the theorizing activity can aid the design of new artifacts [23].

Most recently, Hevner et al. [15] proposed a framework for IS research extending the earlier March and Smith framework. As depicted in Figure 3, this framework defines the environment as a problem space in which the phenomenon of interest resides, and defines the knowledge base as a composition of IS research foundations and methodologies. The center of this framework is the IS research that applies appropriate foundations and methodologies from the knowledge base according to business needs from the environment. Hevner et al. argue that better IS research can be achieved through the combination and integration of two research paradigms—behavioral science and design science. Such integrated IS research is conducted in two phases—develop theories/build artifacts and justify theories/evaluate artifacts. By using multiple research methods, the justifying/evaluating activities can result in better refinement and reassessment of the theory or artifact. Again, different methods may be appropriate to accomplish one to many of the research strategies or activities.

In summary, these three multimethodological, multiparadigm IS research frameworks are actually presenting similar concepts from different perspectives. In this paper, we combine these frameworks into an integrated view of IS research, and explain in detail the interactions among multiple research methods and activities.

**IS Research Activities and Methods**

The four major activities (build, evaluate, theorize, and justify) are all essential parts of IS research. In this paper, we are particularly interested in the interactions between

![Figure 3. Hevner et al.'s IS Research Framework [15]. Reprinted with permission.](image-url)
the theory testing/justifying activity and the artifact evaluation activity and the con-
sequential benefits to IS research. Therefore, in the following sections, we focus on
introducing some of the general research methods used for these two activities. Most
of these methods can be used for both theory testing/justifying and system (artifact)
evaluation, as illustrated in both Nunamaker et al.’s [33] and Hevner et al.’s [15]
frameworks.

Theory Testing/Justifying Activities

From the behavioral science perspective, IS research is viewed to be a branch of the
social sciences [14]. Growing from behavioral science research methods, the primary
goal of the behavioral science paradigm is to develop, evaluate, and validate the truth
of hypotheses and theories developed to inform the design of IS. Theorizing and
justifying activities in the behavioral science paradigm include methods such as case
studies, experiments, and field studies. Experimentation is the predominant method
for theory testing. An experiment is a controlled experience of participation in a de-
dsigned event [2]. As such, the experiment has many characteristics that distinguish it
from other evaluation methods. A classical experiment involves a representative sample
of participants, randomly assigned to multiple treatment groups [2]. The treatment
groups are then measured and compared on the outcome variable(s) of interest to
evaluate theory. McGrath [25] suggests that a well-designed experiment can serve
three underlying research goals: (1) control of extraneous influences, (2) control (i.e.,
manipulation) of independent variable(s), and (3) control of the setting. The signifi-
cant degree of control inherent in experiments allows the discovery of causal rela-
tionships—the critical component of theories used to explain and predict phenomena.
Therefore, the experiment method is a common tool for theory testing.

Although there are innumerable permutations of experiments, McGrath [25] iden-
tifies three broad categories—laboratory experiments, experimental simulations, and
field experiments. These categories differ in their level of precision of measurement
of behavior (which includes control of external influences), realism of the context,
and generalizability of the findings.

Laboratory experiments are conducted in deliberately controlled settings in which
all variables except those being manipulated are ostensibly held constant. In experi-
mental simulations, an attempt is made to create a realism of content in the lab, an
unreal context. The research advantage of both is maximizing the precision of the
measurement of behavior (independent variable) [25]. In contrast, field experiments
attempt to impose the controls of a laboratory experiment in real settings (i.e., in the
field). As a consequence, field experiments trade some level of precision in measure-
ment for realism of the context and resultant participant behaviors.

System Evaluation Activities

From a design science perspective, the crucial goal of IS research is to evaluate the
quality or utility of the artifact—the information system. Such an evaluation requires
the definition of appropriate metrics or attributes of utility. As discussed in Hevner et al. [15], the utility of an artifact refers to the quality or effectiveness of that artifact. It is assessed by questions such as “Does it work?” “Is it an improvement?” or “How well does it work?” To assess the utility of an artifact, researchers develop metrics measuring the constructs of interest, such as functionality, completeness, consistency, accuracy, performance, reliability, usability, fit with the organization, and other relevant quality attributes [15].

In addition to assessing the current system utility, evaluation activities aim to provide information for improving the quality and utility of the entire system. Nielsen [31] described several common methods for evaluating system utility, such as heuristic evaluation, thinking aloud, questionnaires, and interviews. Other methods include experiments and methods that are classified under the observation method described in Nunamaker et al.’s IS research framework. This observation of the actual use of the system can be either from an objective perspective, such as system log or thinking aloud, or from a subjective perspective, such as users’ self-reports in a questionnaire. Surveys (or questionnaires) and interviews are among the two most popular methods for investigating users’ subjective preferences and attitudes regarding the utility of a system [31]. Both methods are easy to implement, as well as very useful for investigating which system features users particularly like or dislike. Questionnaires require users to answer a static set of questions about the system. In contrast, interviews are conducted by an interviewer who asks the users questions about the system and records the users’ answers. Although interviews are usually more flexible than questionnaires, conducting interviews is much more labor intensive than administering questionnaires. Most objective evaluation methods suffer from this problem. Since users can complete questionnaires independently, questionnaires can be administrated to several users simultaneously, saving time and other resources. Therefore, although utility can be evaluated in many different ways, questionnaires will often be the most practical and effective when conducting large-scale studies.

Examples of the Benefits of a Multimethodological Approach

Research has shown that only a relatively small proportion of IS research studies use multiple methods [28]. Only eight examples of multimethodological research studies are summarized in Mingers [27], categorized by how different methods were combined (e.g., employed in sequence or carried out in parallel), and even fewer studies have described explicitly the benefits of such a multimethodological approach. In Table 1, we summarize the benefits identified by three of them.

Beyond these examples from the IS research field, Lee [20] used a criminal justice research study as an example to illustrate the benefits of combining the positivist research approach and the interpretive research approach. The proposed benefits were that the interpretive understanding might explain the disconfirmations of theoretical predictions in controlled empirical testing, and that the empirical and logical rigors of positivism might aid interpretive researchers in choosing among competing interpretive understandings.
<table>
<thead>
<tr>
<th>Study (brief description)</th>
<th>Methods applied</th>
<th>Benefits reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaplan and Duchon [18]. The results of introducing computer technology in a medical laboratory.</td>
<td>Open-ended interview, observation, participant observation, survey (containing closed-ended and open-ended items).</td>
<td>Identified inconsistency of results from different methods. Triangulation of data alerted for potential analytical errors and omissions. Led to new insights and modes of analysis.</td>
</tr>
<tr>
<td>Markus [24]. How and why managers use electronic mail.</td>
<td>Positivist approach: survey. Interpretive approach: inductive and interpretive analysis on interview and written comments.</td>
<td>Triangulated evidence from multiple sources explained the disconfirmations of theoretical predictions.</td>
</tr>
<tr>
<td>Trauth and Jessup [36]. Use group support systems to support employee discussions about gender equity in a university.</td>
<td>Positivist content coding and interpretive open coding on the same set of data from a survey.</td>
<td>The interpretive analysis provided a different understanding of the same evidence and new information not found in the positivist analysis of the group discussions.</td>
</tr>
</tbody>
</table>
Lee’s [20] example exemplifies two general advantages of a multimethodological approach: (1) combined results provide full understanding of the problem and (2) triangulated results provide validation of the results and interpretations. However, we believe that the benefits of the multimethodological approach go beyond these two. Particularly, there can be complex interactions among different methods across different paradigms. We describe such interactions in the next section and illustrate them with the A99 study described in the fourth section.

Interactions Between System Evaluation and Theory Testing

Since the aforementioned research frameworks present similar concepts from different perspectives, we combine them into an integrated view of IS research as illustrated in Figure 4. This integrated view shows that interactions between IS research activities from these two paradigms may complement each other, achieving a more comprehensive understanding of the problem. We hope that by combining the three aforementioned frameworks, and explicitly explaining the interactions between research activities, the proposed integrated framework will provide more robust and explanatory guidance for IS research studies.

As shown in Figure 4, the first type of interaction occurs between the “build” and “theorize” research activities. To reiterate March and Smith [23], building artifacts creates an object for or provides proof of the claims of behavioral science research, whereas the theories found by behavioral science researchers can aid in the design of new artifacts. This type of interaction is somewhat similar to the interaction between engineering and science described in Shaw [35]. For example, in the history of civil engineering [35], people were able to build or craft structures before they understood the theories explaining why those structures worked. The existence of these crafted structures later enabled scientists to discover the underlying explanatory principles. These theories, in turn, enabled the development of standard methods for building structures with predictable and repeatable results. Finally, the new structures provided applied support of the theories. In this way, building artifacts and theorizing can be complementary activities. Every working system embodies certain lawlike principles that explain how the systems work.

In addition, we believe there are at least two other types of interactions that occur between system evaluation and theory testing activities. As illustrated in Figure 4, these additional interactions have the potential to increase the validity of IS research by leading to better theories and better artifacts. In interaction 1, the result–result interaction, the results from one research method used in system evaluation activities interact with the interpretation of results from another method used in theory testing research activities, and vice versa. The result–result interaction may result in either better understanding of the theory, better improvement of the system design, or both. For example, results from a system evaluation survey may give researchers insight into system limitations and thus aid them in interpreting nonsignificant results from a theory testing experiment. As such, the result–result interaction is the commonly recognized benefit of multimethodological research [15, 23, 33].
In interaction 2, the result–design interaction, the results (findings) from a research method used in system evaluation activities interact with the research design of another method used in theory testing activities, and vice versa. Examples of designing methods for system evaluation activities include the selection of measures for evaluation or the selection of evaluation tools (e.g., a survey questionnaire or system log). In contrast, examples of designing methods for theory testing activities include the manipulation of independent variables and the control of experimental settings.

Like the result–result interaction, the result–design interaction is potentially bidirectional. First, the results of a system evaluation method may interact with the design of a theory testing method. For example, results from a system evaluation survey (a design science method) deployed in conjunction with a theory testing experiment (a behavioral science method) may help researchers discover that the actual use of a system varied significantly from the expected use (e.g., not all system functionalities were utilized) due to unforeseen conditions (e.g., the experimental task did not properly motivate the expected use). As a consequence, this discovery can facilitate the identification of hidden limitations in the design of a theory testing experiment (e.g., lack of real differences between the system functionalities used by different treat-
ments groups) and lead to the improvement of the experimental design (e.g., improved task design to motivate expected/intended use). This occurred in A99 Study 2.

In the second form of result–design interaction, the results of a theory testing method may interact with the design of a system evaluation method. In this situation, the results from a theory testing experiment (a behavioral science method) employed in conjunction with system evaluation surveys (a design science method) may provide evidence (e.g., experimental results that contradict survey results) that additional evaluation measures (e.g., survey questions) or tools (e.g., system logs) are needed to capture the actual use or users’ perceptions of the system. Thus the result–design interaction may assure that the theory testing effort is not confounded by the unexpected use of the system in the field; and therefore, the result–design interaction between methods across research paradigms may ultimately improve IS research.

These two types of interactions between the research methods within system evaluation activities and the research methods within theory testing activities have certain implications to IS research. As Hevner and his colleagues stated, “truth and utility are inseparable. . . . An artifact may have utility because of some yet undiscovered truth. A theory may not yet be developed to the point where its truth can be incorporated into design” [15, p. 80]. When testing a theory for understanding the phenomena surrounding an IT artifact, one is actually evaluating the use of that IT artifact [29]. If one is not able to control the use, one should observe/measure the actual use to make sure that unexpected variation of the actual use does not confound the results. On the other hand, in order to understand the use, one should also understand the system and its limitations. To obtain the most comprehensive understanding of an IT artifact and its related phenomena, an IS research study should include both system evaluation activities and theory testing activities. The two types of interactions will help ensure that the findings are not confounded or biased; and presumably result in better IS theories and/or better IS.

With that in mind, we are not proposing that a multimethodological IS research approach should be relied on to evaluate and validate the use and design of methods. As any researcher knows, careful research design and selection of appropriate methods are critical to conducting any research project. However, the most careful and intelligent research design cannot assure the elimination of all limitations or alternative hypotheses, and sometimes research design flaws occur despite researchers’ best efforts. Furthermore, because the actual use of an IT artifact is often difficult to control, unforeseeable research design limitations may exist. In these cases, the result–design interaction provides an additional level of validation for IS research, enhancing the research quality.

The Agent99 Trainer Study: Computer-Based Training for Deception Detection

As mentioned in the first section, we formed our arguments about the interactions among different research methods from the A99 study. To provide an example
demonstrating the two types of interactions and their benefits to IS research, we de-
scribe the A99 study in this section, including the background, system development,
research design, results, and, most importantly, the interactions between different
research methods in the cross-paradigm research activities, system evaluation, and
theory testing. Additional detail regarding the A99 study is presented in George et
al. [13].

Background

The A99 study was designed to test a computer-based training system for training
people to detect deception [12]. Deception is a complex human behavior. The high
failure rate of deception detection (DD) by humans renders the task of training people
to detect deception challenging [9, 26]. Furthermore, the shortage of qualified DD
instructors causes high costs for traditional classroom training [7]. Therefore, there
were three major goals of this study: (1) to build a computer-based training system to
implement DD training programs as an alternative to traditional classroom training,
(2) to find out how this computer-based DD training system worked as compared to
traditional classroom training, and (3) to determine why this DD training system
worked or did not work for its target audience. In other words, the phenomenon of
interest in the A99 study was the effectiveness of training in the domain/context of
deception detection, with the support of IT (see [13] for details). To accomplish the
three major goals, we conducted the study based on Nunamaker et al.’s multimeth-
ological IS research framework. In particular, we iteratively developed system pro-
totypes, evaluated the system using questionnaires/surveys, and tested the theoretical
hypotheses using field experiments. A summary of the A99 study is listed in Table 2.

The first version of the training program was developed in 2001 and the prototype
of the system (named Agent99 Trainer) was evaluated in 2002. Field experiments
were conducted to test our hypotheses and questionnaires were used to test the utility
of this prototype [4, 6, 12]. However, the results from the 2002 experiments did not
support all of our hypotheses. Fortunately, the results of the system evaluation ques-
tionnaires helped us interpret the results of this theory testing experiment, and we
realized that a number of problems with the system implementation might have con-
tributed to the insignificant results of the experiment. Therefore, several improve-
ments were made to the system and new hypotheses were proposed after the 2002
evaluation. The improved system and the hypotheses were evaluated in 2003, again
using a combination of different research methods including field experiments and
questionnaires [7, 11]. The results from the 2003 system evaluation not only helped
interpreting the results of the theory testing experiment but also revealed unforeseen
limitations in the experimental design, highlighting needed revisions for the next
research iteration. These iterative interactions between different research methods
clearly demonstrate the complementary nature of the two types of interactions de-
scribed in the third section.

In the following sections, we chronologically describe the system development and
the consequent studies conducted.
Table 2. Agent99 Trainer (A99) Research Studies

<table>
<thead>
<tr>
<th>Phases</th>
<th>Version</th>
<th>2002 First prototype (designed in 2001)</th>
<th>2003 Second prototype (designed after 2002 study)</th>
</tr>
</thead>
<tbody>
<tr>
<td>System features</td>
<td>System</td>
<td>Two modules.</td>
<td>Professional lecture and example videos with high video/audio qualities.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– View examples: real life examples and feedback.</td>
<td>• Two new modules: search tools (e.g., Ask A Question) and assessment tools (e.g., Quiz)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Synchronization of lecture video, PowerPoint slides, and notes.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Navigation button and pull-down menu allow jump between topics.</td>
<td></td>
</tr>
<tr>
<td>Theory testing</td>
<td>Method</td>
<td>Field experiment with one control group and three treatments (lecture only, A99 only, and a combination)</td>
<td>Field experiment with five treatments groups (1. video only; 2. linear A99; 3. A99 + AAQ; 4. A99 + AAQ + content; 5. A99 + AAQ + content + Quiz).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Procedure:</td>
<td>Procedure:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Pretest (DD accuracy and knowledge) + training + posttest (DD accuracy and knowledge)</td>
<td>• Pretest (DD accuracy and knowledge) + training + posttest (DD accuracy and knowledge).</td>
</tr>
</tbody>
</table>

(continues)
Table 2. Continued

<table>
<thead>
<tr>
<th>Phases</th>
<th>Version</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>First prototype (designed in 2001)</td>
<td>Second prototype (designed after 2002 study)</td>
</tr>
<tr>
<td></td>
<td>Results</td>
<td>H1: significant improvement for all three treatment groups with DD knowledge; but no significant improvement with DD accuracy. H2: no significant differences between A99 and the other treatments.</td>
<td>H1: overall significant improvement with DD knowledge and accuracy. H2: treatment 5 is better than 4, treatment 1 is worse than the others; but no significant differences among treatments 2, 3, and 4 for both DD knowledge and accuracy tests.</td>
</tr>
<tr>
<td>System evaluation activity</td>
<td>Method</td>
<td>Questionnaire with closed- and open-ended questions.</td>
<td>Questionnaire with closed- and open-ended questions (more questions added).</td>
</tr>
<tr>
<td>[5, 6]</td>
<td>design</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Results</td>
<td>• Overall positive: easy to use, good learnability, system features preferred include synchronization, self-controlled navigation, view examples. • Identified problems: system problems, such as lack of interaction.</td>
<td>• Overall positive: most system features useful, quiz very helpful. • Identified problems: users had no time and were not forced to use search tool, so treatments 2, 3, and 4 became the same condition.</td>
</tr>
</tbody>
</table>
System Prototype 1 and Studies in 2002

This section describes the 2002 studies on the first version of the A99 training prototype.

System Design

Our review of the DD training literature [4, 9] identified three critical components of DD training—explicit instruction, practice, and feedback. Since DD is complex, a deep understanding of DD cues requires extensive experience and high levels of cognitive processing. Thus, not only is it important to incorporate the principles of a good training program but it is also important to make the training learner-centered by providing navigational facilities to different modules, library of examples with analysis, and assessment examples with immediate feedback and analysis. With these goals in mind, we developed our first prototype of A99, a computer-based training system for DD. This prototype implemented the three critical components of DD training in two modules [22]: Watch Lecture and View Example with Analysis. The Watch Lecture module provided explicit instruction on deception cues through a combination of synchronized media including the expert video lecture, presentation slides, and lecture notes. The instruction specifically focused on five behavioral categories of cues—arousal, emotion, cognitive effort, memory process, and communication tactics. The lecture was segmented into different topics. Navigational buttons and pull-down menus allowed users to maneuver to any topic or example. The View Example with Analysis module provided real-life examples and scenarios on judging the veracity of communications based on the lectures. It also provided immediate feedback and analysis of the deception cues illustrated.

Pilot Studies and Field Experiment for Theory Testing

Experiment Design. Training for DD has been studied for years, however, results of past training are mixed [9]. Not every training program proved to be able to improve humans’ ability of detecting deception. Therefore, the first question to be answered in our research study was: Can training using reliable cues or indicators of deception truly help people better detect deception? On the other hand, previous research showed that computer-based training, especially Web-based training, can provide benefits over the traditional classroom training such as self-paced, repeatable, and cost-effective access to the curriculum [22]. Although these benefits have been demonstrated in certain general training or learning domains, the question that remained unanswered was whether they would hold true in the DD training domain. Therefore, the two hypotheses tested were: (1) learners receiving DD training perform better on both a DD knowledge test and a DD accuracy test than those receiving no training; and (2) learners receiving training through A99 perform better on a DD knowledge test and an DD accuracy test than learners receiving training by an instructor-led,
lecture-based training method. (See [13] for further discussion of the theoretical underpinnings of these hypotheses.)

Field experimentation was chosen as our research method to test these two hypotheses because it preserves ecological validity while introducing a good measure of control and ability to draw causal inferences [25]. Field testing is also a stringent form of testing for training because of the uncontrolled diversity of the participants, the distractions present in the environment and from the physical settings, and so forth.

The experiment was conducted in fall 2002 at a large U.S. Air Force (USAF) facility located in the United States. A total of 115 officers participated as subjects. The DD training was part of their training program. Rather than a single lecture, which has been a common approach in training research, the design consisted of three one-hour training sessions: (1) an introduction to DD (Intro), (2) cues used to detect deception (Cues), and (3) heuristics for decision making that are susceptible to deception (Heuristics). Subjects were assigned to “blocks,” or classes, made up of 16 students, and blocks were randomly assigned to either the control or to one of three treatment groups. The control group received no training, but they completed the same measurement instruments as the experimental subjects. The treatments were different in the second training session. For one treatment, the lecture-only training groups, traditional lectures were used. For the second treatment, the computer-based training groups, the A99 system was used. For the third treatment, the “combination” group, a combination of lecture and A99 was used. All lectures in all treatments were supported with the same PowerPoint presentations and examples. The training sessions began with subjects completing a series of instruments, including a DD knowledge pretest (multiple-choice questions about the concepts of deception) and a DD accuracy pretest (judging the veracity of recorded interview examples), both directed at determining how much of the subject matter the trainees understood entering the session. Subjects were then trained for approximately one hour, except for control subjects, who were given a one-hour break in the interim. Afterward, all subjects completed a DD knowledge posttest, made up of the same questions as the pretest but in a different order, and a DD accuracy posttest, similar to the pretest but consisting of different examples. The experimental design is summarized in Table 3.

To “debug” the A99 and validate the experimental procedure, two pilot studies were conducted prior to the 2002 field experiment. Participants were undergraduate students enrolled in summer classes offered by the management information systems department at a research one university in the Southwest. The procedure followed was the same as in the field experiment, except that the experimental design was simplified. Only one training session, the cues lecture, was tested, and only the DD accuracy tests were used as the pretest and posttest. The examples in these tests were different from those used in the pretests and posttests of the experiment. There were three instead of four treatment groups in Pilot 1—control, A99, and traditional lecture. Since the results of the first pilot study demonstrated that the pretest and posttest were statistically equivalent and that there was no practice effect between the pretest and posttest, only two treatment groups, A99 and traditional lecture, were tested in Pilot 2.
Table 3. Experimental Design of Experiment 1

<table>
<thead>
<tr>
<th>Training sessions</th>
<th>Treatment conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intro</td>
<td>No training</td>
</tr>
<tr>
<td></td>
<td>Lecture only</td>
</tr>
<tr>
<td>Cues</td>
<td>No training</td>
</tr>
<tr>
<td></td>
<td>Lecture only</td>
</tr>
<tr>
<td></td>
<td>A99 only</td>
</tr>
<tr>
<td></td>
<td>Combination</td>
</tr>
<tr>
<td></td>
<td>Pretests</td>
</tr>
<tr>
<td></td>
<td>Pretests</td>
</tr>
<tr>
<td></td>
<td>Pretests</td>
</tr>
<tr>
<td></td>
<td>One-hour training</td>
</tr>
<tr>
<td></td>
<td>(for the control group with no training, there was a one-hour break)</td>
</tr>
<tr>
<td>Heuristics</td>
<td>No training</td>
</tr>
<tr>
<td></td>
<td>Lecture only</td>
</tr>
<tr>
<td></td>
<td>System try-out</td>
</tr>
<tr>
<td></td>
<td>Questionnaire</td>
</tr>
</tbody>
</table>

Results. Two dependent variables were measured in the field experiment: the DD accuracy measured by the number of correct judgments divided by the total number of test cases in each DD accuracy test, and the DD knowledge measured by the number of correct answers divided by the total number of multiple choice questions in each knowledge test. Independent t-tests showed that DD knowledge improvements in the treatment groups differed from the control group for all three sessions (Intro: \( t(113) = -8.921, p < 0.001 \); Cues: \( t(113) = -4.54, p < 0.001 \); Heuristics: \( t(113) = -7.536, p < 0.001 \)). It revealed that the subjects’ DD knowledge was significantly improved in the three training session groups, while the control group did not improve, just as we predicted. However, there were no significant differences between A99 and the other delivery modes on DD knowledge tests. In addition, there was no significant improvement on DD accuracy in all groups, which did not support our prediction (detailed data analysis can be found in [12]).

Therefore, a major question for us at that time was why the results did not support all our predictions. Was it the case that training does not improve DD accuracy? Was it the case that computer-based training does not provide extra benefits beyond traditional classroom learning? Or might the nonsignificant findings attributable to problems with the design of the Agent99 Trainer? Had we only conducted the field experiment, the answers would have been indeterminate. Fortunately, both the pilot studies and the system evaluation questionnaires deployed at the end of the experiment provided us potential explanations for these results. Pilot 2 showed a significant improvement on DD accuracy (A99: pretest 0.4222, posttest 0.6889; lecture: pretest 0.4405, posttest 0.6429; \( F(1, 27) = 32.29, p < 0.01 \) (detailed data analysis can be found in [12]). Given that the experiment procedure and the training contents were virtually the same for the field experiment and the pilot studies, and only the examples used in the DD accuracy pretests and posttests were different, a plausible explanation for the nonimprovement in detection accuracy was the use of problematic pretest or posttest questions. The system evaluation activity, on the other hand, helped us understand why there was no difference between the A99 and the traditional classroom training.
Questionnaires for System Evaluation

A system evaluation questionnaire was employed to test whether the design of the first prototype of the A99 met its objectives. Since users’ subjective attitudes toward the system were the major interest of this study, attributes such as learnability, ease of use, satisfaction, and effectiveness/usefulness were selected to be measured from several different perspectives [6]. The questionnaire was iteratively revised after each pilot study or experiment, but all versions contained both closed-ended items rated on a seven-point Likert scale and open-ended questions. The closed-ended items were obtained either directly from an existing validated measure called the system usability scale [3], or adapted from other existing validated measures such as the computer usability satisfaction questionnaires [21] and perceived usefulness and ease of use [8], or developed by us regarding some specific system features.

The questionnaires were administrated at the end of the pilot studies or at the end of the experiment. For the field experiment, since only two treatment groups used A99 during session 2 in the experiment, the other two groups (control and lecture) were given 30 minutes to use A99 at the end of session 3 before completing the questionnaire (see Table 3). All groups completed the questionnaire online after all other phases of the experiment were completed. Unfortunately, because of a computer programming error, the online responses to the closed-ended questions were miscoded and noninterpretable. Therefore, only the responses to the open-ended questions in the field experiment were analyzed. For the pilot studies, participants in the A99 group completed the questionnaire and the responses for both closed-ended and open-ended questions were analyzed.

Results. For both Pilot 1 and Pilot 2 data, means were calculated on each closed-ended item and on a smaller number of composite measures extracted by a principal components factor analysis that represented the major usability attributes. The results from both Pilot 1 and 2 were highly positive, with all means less than or close to 3 ("completely agree"—"slightly agree"; see [6] for more details). This indicated that most of the students agreed that A99 had good learnability, was easy to use, and so on. Both the delivery method (structured and synchronized multimedia) and the training contents (especially examples and analyses) used in A99 were perceived as very effective for DD training. One of the major features of the A99 design—the self-paced learning method—was also evaluated very favorably. However, despite the positive ratings, subjects using A99 for training did not perform better than those receiving traditional training. Why not?

The qualitative data collected from the open-ended items in the questionnaires provided important clues about the users’ actual usage of the system in the experiments, and thus helped answer the above question. Responses on the open-ended questions showed that users liked that the system was easy to use, provided a structured and synchronized multimedia lecture, and afforded user control of their learning pace. Furthermore, responses indicated that the examples and analyses were important to the DD training effectiveness. However, users also described some problems with the system, such as inappropriate navigation, the relatively low audio/video quality caused...
by either the file compression necessary for Web-based delivery or nonprofessional video production in some of the older recorded deception examples. Furthermore, the responses indicated that students viewed interaction with instructors as an important feature of good training, something that did not exist in the current system. These negative comments indicated that perhaps the Agent99 group did not outperform the classroom training group in our experiments because of system design problems, weakness, and lack of important features.

In short, the results from the system evaluation activities yielded invaluable information that contextualized the results from the theory testing activities and informed interpretations of the results. The combined research methods demonstrate the benefits of the result–result interaction proposed in the third section.

System Prototype 2 and Studies in 2003

Based on the results and lessons learned from the 2002 experiment and system evaluation, we not only redesigned the A99 system but also redesigned the experiment and refined the system evaluation questionnaire.

System Design Improvement

The second prototype of A99 was designed to overcome the shortcomings of the first prototype revealed through the prior experiments and the system evaluation questionnaires. To deal with the bandwidth restriction problem and the fact that the environment for intended users of this system, Air Force officers, posed problems regarding networking security, the delivery media of the second prototype was changed from the Web to CD-ROMs. The Watch Lecture and View Example with Analysis modules were integrated into a single Virtual Classroom module with more fine-tuned synchronization between the lecture video, presentation slides, and lecture notes. The interface was revamped based on the user comments from the 2002 studies. Examples were embedded in the lecture video, while the pull-down menu was replaced by an outline-type navigation menu facilitating better maneuverability to choose any topic. Two new modules called Search Tools and Assessment Tool were designed to provide better interaction and assessment capabilities. The Search Tool included components such as Key Word Search and Ask A Question (AAQ) to facilitate user-driven information delivery. The search query retrieved relevant video segments in the lecture video. The assessment tool included pop-up quizzes at certain break points to facilitate self-assessment and immediate feedback through correct answers. Thus, the new design improved the earlier prototype based on the results of the 2002 studies.

Field Experiment for Theory Testing

Because the results from the 2002 studies did not fully support our hypotheses, especially the hypothesis that the A99 can provide better learning achievement than traditional training, a field experiment was conducted again at the same USAF facility in
2003 but with several design changes. First, for the classroom training treatment, video lectures were used instead of live instructors to standardize the presentation order and content. The video production quality was improved to produce professional audio and video. Second, to test whether the last year’s failure was caused by the partial implementation of A99 and whether the newly added functionalities could improve the training, the other treatment groups were designed to use three different versions of the A99. Finally, although both the literature and the results from the 2002 system evaluation indicated that examples and analyses were important for making DD training effective, we wanted to know explicitly whether more examples would make the training more effective. Therefore, the participants, 180 Air Force officers, were randomly assigned to the following five treatment groups, with each added functionality expected to enhance participants’ performance [11].

- **Treatment 1: Video Only.** Users watched a video of an expert talking about deception, with PowerPoint slides and examples cut in the video. The presentation order in this video was predetermined by the instructor.

- **Treatment 2: Linear A99.** Lecture outline, expert video, PowerPoint slides, and lecture notes were shown on the A99 interface. PowerPoint slides and lecture notes were synchronized with the lecture video. However, users could not click on the lecture outline to change topics. They still had to follow the predetermined instruction pace and sequence.

- **Treatment 3: A99 + AAQ.** Users could control their own learning pace by clicking on the lecture outline. They could also look for specific topics using key word search or AAQ.

- **Treatment 4: A99 + AAQ + Content.** Links to more examples of DD were provided at the bottom of the lecture outline.

- **Treatment 5: A99 + AAQ + Content + Quizzes.** This treatment deployed the complete functional implementation of A99, including pop-up quizzes. In addition, more examples of DD were also provided in this treatment.

The experimental procedure followed the same pretest training–posttest procedure as in the 2002 experiment, except that there were only two training sessions in this experiment. The session on “heuristics” was removed and its contents were redistributed to the remaining two lectures. The two sessions were each approximately one and a half hours in duration. The DD accuracy pretest and posttest examples, however, were modified carefully, since the previous results indicated that there could be some problems with the test questions. The experimental design is summarized in Table 4.

**Results.** For both the introductory and the cues training sessions, there was an overall significant improvement on both the DD knowledge test (Intro: \( t(176) = 17.53, p < 0.001 \); Cues: \( t(176) = 5.82, p < 0.001 \)) and the DD accuracy test (Intro: \( t(174) = 3.88, p < 0.001 \); Cues: \( t(175) = 9.95, p < 0.001 \)) [11]. Therefore, our prediction that training can improve humans’ DD capability was supported. An analysis of variance (ANOVA) test with planned contrast analysis for knowledge pretests and posttests
Table 4. Experimental Design of Experiment 2

<table>
<thead>
<tr>
<th>Treatment conditions</th>
<th>Video</th>
<th>Linear</th>
<th>A99 + AAQ + content</th>
<th>A99 + AAQ + content</th>
<th>A99 + AAQ + content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training sessions</td>
<td>Intro</td>
<td>Pretests</td>
<td>Instruction</td>
<td>Pretests</td>
<td>Posttests</td>
</tr>
<tr>
<td></td>
<td>Cues</td>
<td>Pretests</td>
<td>Instruction</td>
<td>Pretests</td>
<td>Posttests</td>
</tr>
</tbody>
</table>
found near-significant differences between treatments 4 and 5 for both the Intro \( (p = 0.090, \text{ one-tailed}) \) and the Cues \( (p = 0.051, \text{ one-tailed}) \) training sessions, suggesting that quizzes did improve the learning of DD knowledge. For DD accuracy tests, the contrast analysis discovered that for the Intro session alone there was a significant difference between treatments 4 and 5 \( (p = 0.033, \text{ one-tailed}) \), as well as between treatment 1 and the other four groups \( (p = 0.035, \text{ one-tailed}) \). It indicated that the A99 system did improve DD accuracy compared to watching a single video. The results also indicated that the pop-up quiz was indeed an important feature contributing to improve the detection accuracy. However, conditions 2, 3, and 4 did not differ significantly from one another even though these conditions provided very different functions to users. Did it mean that, even though both the interactivity and the examples were the functions that the 2002 systems evaluation showed users actually desired, the new functions (the search tools and more examples) did not contribute at all to the training effectiveness? Again, the information gathered by the system evaluation questionnaires provided us insights into what had happened in the experiment.

**Questionnaires for System Evaluation**

Because the system was modified considerably and many new functions were added, the system evaluation questionnaires were deployed again to evaluate the utility of this new version of the system. With the experiences from the 2002 studies, we also expected that the results from the system evaluation would aid in interpreting the results from the field experiment. Because the Intro lecture had less content than the Cues lecture, the questionnaires were given after the posttest at the end of the first session of the experiment to prevent participants from becoming fatigued. The questionnaire was modified according to the system changes, although we still focused on similar metrics such as ease of use, satisfaction, effectiveness/usefulness, and so on. In this version of the questionnaire, there were 29 closed-ended questions rated on a seven-point Likert-type scale. Among these questions, seven were about certain system functions that were only available in certain treatment groups. Therefore, an eighth choice, “N/A” (not applicable), was added to deal with this situation. Six open-ended questions allowed users to describe their experiences with the system.

**Results.** In brief, the responses were positive across all conditions, with most means close to or greater than 5 (“slightly agree”→“completely agree”; see [7] for more details). This indicated that the participants agreed that the A99 system had good utility even with just partial functions. Most of the items related to specific features of A99 (e.g., synchronized multimedia presentation and the self-paced learning method) received positive responses and were viewed as effective in helping users learn to detect deception. The newly added function, pop-up quizzes, was also rated as very helpful in supporting learning, which strongly supported the findings from the experiment. However, for the three items measuring users’ attitudes toward the search tools, almost all users selected “N/A,” which meant that they either did not know the functions were available to them, or they knew the functions existed but never used...
them. What happened? After carefully analyzing the users’ responses to the open-ended questions, we realized that most users actually had no time to use these new functions at all, because of the time constraint on the lecture and because they were not forced to use these functions (not as the pop-up quizzes, which were mandatory to be taken during the training). This finding made us realize that even though we designed the conditions 2, 3, and 4 to be different, in the real experiment the conditions were actually the same. No wonder we could not find the differences that we expected. The design of our research method was somewhat flawed.

In addition, five composite variables that represent the major system utility metrics were extracted by a principle component factor analysis from 22 closed-ended items that were applicable to all five conditions—perceived learning effectiveness, general satisfaction, audio/video quality satisfaction, ease of use/learning to use, and comparison with traditional classroom learning. Results from ANOVA tests with planned contrast analysis revealed that the perceived learning effectiveness in all A99 groups (treatments 2, 3, 4, and 5) was significantly better than in the video-only group ($t = 1.697, p = 0.046$, one-tailed). Both the perceived learning effectiveness ($t = 1.730, p = 0.043$, one-tailed) and the general satisfaction ($t = 2.366, p = 0.010$, one-tailed) in treatment 5 were significantly better than the other A99 groups with partial functions (treatments 2, 3, and 4). All the other variables were statistically equal across all conditions (see [7] for details). These findings were consistent with the results from the experiment, and reconfirmed our conclusion: the mistakes in the experimental design, especially the time arrangement and the motivation issues, resulted in the unexpected experimental results.

In summary, in the second phase of the A99 study, the results from the system evaluation activities not only provided important context for interpreting the results from the theory testing activities but also identified unforeseen limitations in the theory testing method design, demonstrating benefits from both the result–result interaction and the result–design interaction proposed in the third section.

Discussion

As described in the above section, our three-year, iterative, multimethodological A99 study demonstrated the existence of two types of interaction between research methods across two different research activities, as well as their benefits to the IS research. As illustrated in Figure 4, the result–result interaction potentially results in either better understanding of the theory, better improvement of the system design, or both. For example, in the A99 study, the 2002 system evaluation results helped us understand why there were no differences between the treatment groups in the experiment: there were a number of problems with the system implementation. The result–design interaction, on the other hand, may result in better research design. For example, the 2003 system evaluation results revealed that users were neither given enough time to complete the training nor forced to use specific functions; therefore, three “different” treatments became identical in the experiment. This finding not only accurately explained why there was no difference in the results between these treat-
ments but, more importantly, caused modification of the experiment design. Without the system evaluation results to identify the experimental design problems, the same experiment might have been conducted repeatedly without finding the real truth.

Therefore, the two types of interactions, the result–result interaction and the result–design interaction, in the A99 study prevented the variation of the actual use of the system from the expected use from confounding the study of the factors influencing system success, presumably resulting in better, unbiased theories. At the same time, the better understanding of the system helped us pay attention to and interpret the actual use of the system, which resulted in a better artifact—the information system.

Some may argue that although the two types of interactions are beneficial to IS research, combining system evaluation and theory testing activities requires too much time and effort to make it worth trying. It is true that the A99 study has already lasted for three years and is still continuing. However, one should realize that much of the research time was required for building the A99 system. The system evaluation and theory testing activities and methods were actually deployed at the same time. We understand that not all researchers are interested in or have the skills for building IS. Furthermore, in many instances, researchers just want to investigate the phenomena surrounding an existing system. Similarly, researchers may only be interested in building an IS based on some existing theories. Therefore, we do not suggest that every IS research study includes all four research activities—build, evaluate, theorize, and justify. However, we do suggest that to achieve the most comprehensive understanding about an IT artifact and its related phenomena, an IS research study should include both system evaluation activities and theory testing activities. From our experience, these two activities can be integrated easily by using appropriate research methods. The benefits from the interactions between the methods will outweigh the efforts of deploying them. For design science researchers, adding theory testing activity will help them understand the system and its limitations and thus better understand the actual use of the system. Similarly, by adding system evaluation activities, behavioral science researchers can observe/measure the actual use of a system to make sure that the findings are not confounded or biased, even if they may not be able to predict or control the use.

Again, technology and behavior are inseparable. We should recognize that IS research should be conducted in an integrated way, consisting of multiple research activities and methods. The current two paradigms are two sides of this integrated approach. Although each side can be conducted independently (and definitely certain methods will fit best for certain research problems) and may contribute equally to the IS research knowledge base, combining them gives IS researchers a more comprehensive understanding of the problem. In addition to reemphasizing the importance of an integrated approach, this paper contributes to the IS literature by combining three earlier multimethodological IS research frameworks together into one, more robust and explanatory framework. The resulting framework not only explains what research activities and methods can be used in a multimethodological IS research approach but also indicate how these research activities can interact with each other.
Furthermore, we suggest that the benefits from these interactions will generally outweigh the effort of deploying them, and will result in better IS theories or better IS. However, one limitation of the work presented in this paper is that only the interactions between two specific research methods—field experiment and questionnaires—were discussed in a specific research context (the development and investigation of A99). Therefore, the generalizability of our work is limited. Nonetheless, at this time, this limitation is unavoidable since such an iterative, multimethodological study is time-intensive, mitigating the possibility of conducting multiple cases simultaneously. We hope that other researchers will follow our suggestions in conducting their own research to reap the benefits from the interactions between research methods. We anticipate that these new cases will confirm or extend the types of interactions identified in this paper. But, most importantly, we hope to see that such interactions result in better IS research.

Acknowledgments: This research was supported by funding from the U.S. Air Force Office of Scientific Research under the U.S. Department of Defense University Research Initiative (Grant No. F49620–01–1-0394). The views, opinions, and findings in this paper are those of the authors and should not be construed as an official Department of Defense position, policy, or decision. Parts of this paper have been published as conference proceedings at ISI 2003, ISI 2004, HICSS-37, AMCIS 2003, and ICIS 2004. An abstract of this paper was originally published in the Proceedings of the 2004 International Conference on Information Systems (ICIS 2004).

Note

1. The term interaction used in the context of this paper means the kind of action that occurs as two or more objects have an effect upon one another. It needs to be distinguished from the meaning typically used in data analysis that refers to the statistical variations among means at different levels of two causal constructs.

References


