Making and Evaluating Participant Choice in Experimental Research on Information Technology: A Framework and Assessment

Nancy K. Lankton
Marshall University, lankton@marshall.edu

Joan Luft
Department of Accounting and Information Systems, Michigan State University

Follow this and additional works at: https://aisel.aisnet.org/cais

Recommended Citation
DOI: 10.17705/1CAIS.03511
Available at: https://aisel.aisnet.org/cais/vol35/iss1/11
Making and Evaluating Participant Choice in Experimental Research on Information Technology: A Framework and Assessment

Nancy Lankton
Division of Accountancy and Legal Environment, Marshall University
lankton@marshall.edu

Joan Luft
Department of Accounting and Information Systems, Michigan State University

Abstract:
Evaluations of participant samples for experiments in information systems research often appear to be informal and intuitive. Appropriate participant choice becomes a more salient issue as the population of information technology professionals and users grows increasingly diverse, and the distribution of relevant characteristics in participant samples such as age, gender, nationality, and experience can often be unrepresentative of the characteristics’ distribution in target populations. In this paper, we present a framework based on widely accepted standards for evaluating participant choice and providing rationale that the choice is appropriate. Using a step-by-step approach, we compare current practice in experimental studies from top information systems journals to this framework. Based on this comparison, we recommend how to improve the treatment of participant choice when evaluating the validity of study inferences and how to discuss the tradeoffs involved in choosing participant samples.

Keywords: Experimental Research, Sample Choice, Individual Characteristics, Experimental Participant.
I. INTRODUCTION

As the population of information system (IS) designers and users grows increasingly diverse—from leading-edge scientists to children, third-world villagers, and executives—questions about the validity of results obtained from using limited samples become increasingly salient. This is particularly the case for experimental research in IS, which often uses convenience samples of employees from a particular organization or students from a particular university, who are arguably often not representative of a study’s target population. That is, individuals in the sample differ from the study’s target population with respect to characteristics such as age, gender, nationality, or experience that (a) are not the variables of primary interest in the study and (b) could pose threats to the validity of the study’s inferences about the variables of primary interest. Hereafter, we refer to these characteristics as “sample characteristics”.

A recent paper (Compeau, Marcolin, Kelley, & Higgins, 2012) expresses strong concerns about how researchers are treating sample choices in IS research. Examining studies that use individual-level human subject data in two IS journals, Compeau et al. (2012) found that only a minority of these studies provided any justification for their sample choice or discussion of the resulting limitations. Moreover, in the authors’ view, when justifications were provided, they were often inadequate. As such, they recommend that future research treat sample-choice issues much more carefully by identifying specific similarities between characteristics of the sample and “important population characteristics” and identifying specific limitations to generalizability that result from any dissimilarities between sample group and target population.

IS researchers are thus faced with the questions: how do we form and justify beliefs about which characteristics are important? If there are dissimilarities between sample and population, which of them limit the inferences that can be drawn from the research, and how do they limit these inferences? In this paper, we provide a framework for answering these questions by integrating best-practice recommendations from multiple sources. This framework includes a step-by-step approach that considers the statistical properties of the sample characteristics, threats to validity resulting from these properties, and judgments about which sample and empirical-modeling choices will result in the best tradeoffs—that is, which choices will reduce some important threats to validity with no increase, or an acceptably small increase, in other threats.

We find that current practice in IS experimental research rarely discusses experiment participant choices in terms of these tradeoffs. Researchers often omit the statistical tests that could help determine the appropriateness of a study’s participant choices and the magnitude of the remaining threats to validity, even when they have collected the sample characteristic measures that would enable them to perform these tests (see Section 5 below). Discussions about how participant choice influences the validity of a study’s inferences are perhaps too often limited to formulaic remarks such as “Results should be replicated with other groups” or “The use of student participants is justified by their use in other studies”¹. Additionally, researchers often do not identify specific threats to validity that they were not able to eliminate in their study. Nor do they explain why their study’s design represents a good tradeoff between reducing some threats and failing to reduce others.

The paper proceeds as follows. In Section 2, we describe our data collection methods. In Section 3, we show, consistent with Compeau et al. (2012), that participant samples in IS research are not prima facie representative of target populations. In Sections 4 to 8, because samples that are imperfectly representative do not always pose unacceptable threats to a study’s validity, we provide a five-step approach for addressing validity issues related to imperfectly representative samples: (1) identify and measure potentially influential sample characteristics, (2) examine the variability of these sample characteristics, (3) test or otherwise assess statistical properties of the sample characteristics to determine what specific threats they could pose to the validity of the study’s inferences, (4)

¹ With this paper, we do not intend to contribute to the extensive literature that debates whether and how students as participants are different from various subpopulations of nonstudents (e.g., Rosenthal & Rosnow, 2009; Sears, 1986; Gordon, Slade, & Schmitt, 1986; Henrich, Heine, & Norenzayan, 2010). Rather, we provide criteria for determining when and how these differences—or any other differences between participant samples and target populations—matter to IS research. Differences undoubtedly exist—for example, students have less-formulated senses of self and stronger tendencies to comply with authority than the population at large (Sears, 1986). These differences clearly matter to social psychology and behavioral science research that measures such characteristics in the population at large and in subpopulations. But, often, these differences do not matter to the questions of interest to IS research. For example, individuals with more- and less-formulated senses of self may be similarly influenced by graphic interfaces.
report these specific threats, and (5) explain the trade-offs among them. For each step, we also examine current practice in IS experimental research and recommend improvements.

II. DATA COLLECTION

To examine current practice in IS research, we reviewed papers published from 2000-2012 in four premier IS journals, MIS Quarterly (MISQ), Information Systems Research (ISR), Journal of Management Information Systems (JMIS), and Journal of the Association for Information Systems (JAIS). These journals are in the Senior Scholars’ Basket of Six, and were recently ranked as the top four information systems journals (Lowry et al., 2013), which makes them primary sources of “best practice” observations. Our final sample includes 184 experimental studies, including 53 from MISQ, 46 from ISR, 57 from JMIS, and 28 from JAIS (Appendix and Table 1). These studies account for 11 percent of the total research published in these journals over the 13-year period (Table 1). To collect the studies, we first defined “experimental studies” as those involving human subjects and in which at least one independent variable (IV) was manipulated and randomly assigned to participants to test one or more hypotheses. We excluded mathematical-simulation experiments, surveys, and usability studies that primarily validate the functionality of new software rather than test theory-based hypotheses. Next, the first author examined all papers published in the four journals during this period by reading abstracts, and skimming/reading the paper if needed. Then, the second author and a graduate assistant also examined papers in these journals using this same method. The two authors discussed those papers that only one identified as experimental, or that were borderline cases2, to determine whether to include them in the study.

Table 1: Experimental Research Studies Published 2000-2012

<table>
<thead>
<tr>
<th>Year</th>
<th>MISQ papers</th>
<th>ISR papers</th>
<th>JMIS papers</th>
<th>JAIS papers</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exp.</td>
<td>Total</td>
<td>Exp.</td>
<td>Total</td>
<td>Exp.</td>
</tr>
<tr>
<td>2000</td>
<td>3</td>
<td>23</td>
<td>5</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>2001</td>
<td>2</td>
<td>16</td>
<td>3</td>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>2002</td>
<td>2</td>
<td>17</td>
<td>2</td>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>2003</td>
<td>2</td>
<td>22</td>
<td>3</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>2004</td>
<td>2</td>
<td>24</td>
<td>2</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>2005</td>
<td>2</td>
<td>28</td>
<td>2</td>
<td>21</td>
<td>5</td>
</tr>
<tr>
<td>2006</td>
<td>8</td>
<td>42</td>
<td>7</td>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>2007</td>
<td>3</td>
<td>30</td>
<td>1</td>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>2008</td>
<td>3</td>
<td>34</td>
<td>5</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>2009</td>
<td>6</td>
<td>43</td>
<td>5</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>2010</td>
<td>7</td>
<td>37</td>
<td>3</td>
<td>53</td>
<td>2</td>
</tr>
<tr>
<td>2011</td>
<td>9</td>
<td>50</td>
<td>5</td>
<td>47</td>
<td>5</td>
</tr>
<tr>
<td>2012</td>
<td>4</td>
<td>60</td>
<td>3</td>
<td>74</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>426</td>
<td>46</td>
<td>403</td>
<td>57</td>
</tr>
</tbody>
</table>

For each study, we identified information about the sample, the target population, and the experimental task. We also collected information about any sample characteristics measured and how these sample characteristics were used in the study’s data analysis. The first author and a graduate assistant did this coding. A random check of 20 percent of the coding of sample-characteristic information by the second author yielded no systematic differences in coding outcomes and no differences in conclusions about reporting practice in these experiments. Tables 2 through 6, which form the basis for the following analyses, summarize this information.

III. SAMPLES AND TARGET PARTICIPANTS IN IS EXPERIMENTS

To provide evidence on the apparent lack of representativeness in IS experiment-participant samples, we replicated and extended Compeau et al.’s (2012) findings of questionable matches between sample and target populations. We used a broader and newer sample of research3 than theirs, and we compared participant samples not only with explicitly stated target populations (which, as Compeau et al. (2012) observe, are relatively infrequent) but also with the target populations that were implicit in a study’s research question or in the experimental task or technology employed. For example, if a paper’s research question was how website design influenced consumers’ willingness to transact with an online seller, then, in the absence of further qualifications (e.g., targeting only consumers in certain income or age ranges), we assumed the target population was online consumers. Similarly, if the task was to

---

2 For example, we included some studies in which randomization of participants to treatments was limited (e.g., different treatments could not be assigned to students in the same class section).

3 We examined four journals rather than two. We also began our sample in 2000 rather than 1990 to help exclude possibly outdated practices.
decide whether an organization should invest in particular software, then, in the absence of further qualifications, we assumed the target population was individuals who make software-investment decisions for organizations.

Table 2 overviews the sample and target population choices in the 184 studies we analyzed. The first column describes the participant sample in each study. The second column describes the target population. The third column describes the tasks and technologies used in the experiment (as a check on our judgment of implicit target populations). The fourth column presents the number of studies that include the sample/target population/tasks combination identified in each row.

Data in Table 2 shows that target populations—for example, “software application users needing training” or “organizational members performing group tasks”—were often highly diverse in terms of characteristics that might influence their IS-related behavior, such as age, education, nationality, and personality traits. Often, however, samples were students drawn from a single course or degree program at a single university or employees drawn from a single organization or organizational unit. These samples can be quite homogeneous with respect to some potentially influential characteristics. They can also be diverse with respect to some of these characteristics, but, in many cases, they are likely to lack representative diversity. For example, ages or nationalities in the sample might be quite diverse in both the sample and target population but also quite differently distributed in the two groups.

Thus, in these IS experiments, samples are often not prima facie representative of the target populations with respect to potentially influential sample characteristics; that is, characteristics that might influence the dependent and/or independent variables (DVs and IVs) in the experiment. Although representativeness is a desirable property of samples, we do not simply argue for more representative samples here. Rather, we provide an approach for evaluating threats to validity that arise from imperfect representativeness when full representativeness is not practically attainable. Sometimes it is not attainable—or at least cannot be verified—because the distributions of potentially influential characteristics in the target population are not known with sufficient exactitude. Sometimes, when multiple characteristics are potentially relevant, full representativeness is not attainable because an available sample that is a good match to the population with respect to some characteristics is a poor match with respect to others, and no available sample is an equally good match on all characteristics. Sometimes obtaining a fully representative sample is costly, and researchers may question whether the resulting increase in validity is significant enough to justify the cost. As we argue below, imperfect representativeness—that is, lack of matching on potentially influential characteristic distributions between sample and population—does not always create a threat to valid inference, even when the sample characteristics have significant effects on the variables in the study.

When mismatches between sample and population characteristics do create threats to validity, researchers often face tradeoffs in research design because a choice that reduces one threat can increase another. Researchers then need to make and explain judgments about their sample choice and its consequences for the validity of their research in terms of these tradeoffs. As Compeau et al. (2012) point out, the explanations of sample choice that appear in the literature are often rather perfunctory: they are based on prior practice or simple assertion rather than any more rigorous, theory-based approach. In the following sections, based on widely accepted standards for assessing validity in empirical research (e.g., Shadish, Cook, & Campbell, 2002), we propose a more systematic approach to explaining sample choice and its consequences. We also compare existing practice to these standards and, in the process, often document a substantial gap between current and best-practice treatment of sample-characteristic issues in the IS literature.

IV. STEP ONE: IDENTIFY AND MEASURE POTENTIALLY INFLUENTIAL SAMPLE CHARACTERISTICS

Figure 1 illustrates a five-step approach to identifying and analyzing threats to validity arising from potentially influential sample characteristics that may not be representative of a study’s target population. Dotted-line boxes on the right-hand side of the figure present summary statistics on particularly visible inconsistencies between current and best practice. We present this as an ex post approach in which researchers have already (at least tentatively) selected a sample that they believe is appropriate and now must verify that belief to satisfy themselves and others of their research’s validity. However, the steps in Figure 1 can also be used as part of the ex ante judgment process in selecting a sample while researchers consider the sample’s likely consequences on validity.

The first step in Figure 1 is to identify and measure potentially influential sample characteristics and to explain why these characteristics are potentially influential. Researchers usually cannot identify with certainty a priori all the characteristics of the individuals in the target population and the sample (e.g., specialized training or willingness to take risks) that might influence the variables of interest in their studies. But IS and other social-sciences literature can often provide a basis for identifying a set of potentially influential characteristics. These characteristics should then be measured in the sample.
Table 2: Participants, Tasks, and Target Populations for IS Experiments

<table>
<thead>
<tr>
<th>Participants</th>
<th>Target population</th>
<th>Tasks/technology identified</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergrad students</td>
<td>Stakeholders in the system development process</td>
<td>Reading and developing data models, developing queries, buying software</td>
<td>16</td>
</tr>
<tr>
<td>Group decision-makers using DSS tools</td>
<td>Generating alternatives, choosing among alternatives, completing projects, identifying deception</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Online consumers</td>
<td>Evaluating/buying products/websites, using online agents</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Software application users</td>
<td>Undergoing training, completing skill tests</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><strong>Subtotal studies using undergraduate students</strong></td>
<td></td>
<td></td>
<td>65</td>
</tr>
<tr>
<td>Undergrad and graduate students</td>
<td>Group or individual decision-makers using DSS</td>
<td>Choosing alternatives, developing a business plan, negotiating prices</td>
<td>7</td>
</tr>
<tr>
<td>Stakeholders in the system development process</td>
<td>Reading conceptual models, programming, making project continuance decisions, reusing software</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Online consumers</td>
<td>Browsing websites/products, using mobile devices, participating in virtual worlds and online communities</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Organizational members logging on to systems, using information from various formats/displays</td>
<td>Making personality judgments, interpreting graphical data, assessing security issues</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><strong>Subtotal studies using undergraduate and graduate students</strong></td>
<td></td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>Graduate students</td>
<td>Managers using DSS</td>
<td>Making decisions, choosing alternatives, negotiating</td>
<td>4</td>
</tr>
<tr>
<td>Stakeholders in the system development process</td>
<td>Making technology investment decisions</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Managers using a supply chain management system</td>
<td>Procuring goods</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Electronic market participants</td>
<td>Examining seller information, bidding</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Subtotal studies using graduate students</strong></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Unspe- cified Students</td>
<td>Stakeholders in the system development process</td>
<td>Reading conceptual models, performing systems analysis, making software project decisions, querying</td>
<td>7</td>
</tr>
<tr>
<td>Online consumers</td>
<td>Providing information, browsing websites, evaluating products, examining seller profiles, using virtual reality</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Software application users</td>
<td>Training, completing skill tests</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Organizational members and professionals making decisions, performing group work</td>
<td>Making real-time dynamic decisions and solving problems using a DSS or other collaborative software</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td><strong>Subtotal studies using unspecified students</strong></td>
<td></td>
<td></td>
<td>29</td>
</tr>
<tr>
<td>Students and professionals or other non-student samples</td>
<td>Managers using DSS</td>
<td>Choosing alternatives, decision making, choosing alternatives using a graphical DSS, analyzing deception</td>
<td>5</td>
</tr>
<tr>
<td>Online consumers</td>
<td>Browsing websites, bidding for and evaluating products</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>B2B e-commerce participants</td>
<td>Entering transactions using an exchange technology</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Software application users</td>
<td>Completing skill tests</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Professional and administrative organizational workers</td>
<td>Entering transactions using an exchange technology, examining fear appeals about computer security</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><strong>Subtotal studies using students and professionals or other non-student samples</strong></td>
<td></td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>Professionals or other non-student samples</td>
<td>Online consumers</td>
<td>Searching sites and making purchasing decisions, using e-negotiations, using online agents, reviewing privacy messages, examining seller information, bidding</td>
<td>13</td>
</tr>
<tr>
<td>Professionals using a learning management system</td>
<td>Using system for continuing learning</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Users of various database systems</td>
<td>Using different systems to generate ideas, make decisions, analyze problems, deceive, reach consensus</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Stakeholders in the system development process</td>
<td>Using conceptual models, making software project continuance decisions, giving requirements for a system development project, modifying code</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td><strong>Subtotal studies using professionals or other non-student samples</strong></td>
<td></td>
<td></td>
<td>26</td>
</tr>
<tr>
<td><strong>Total Studies</strong></td>
<td></td>
<td></td>
<td>184</td>
</tr>
</tbody>
</table>
Current Practice

Twenty-one percent of the IS experiments\(^4\) we examined do not report measuring sample characteristics. In studies that did report information on sample characteristics, only 8 percent explain the choice of all reported characteristics, another 21 percent explain the choice of some of the reported characteristics, and the final 71 percent do not explain the choice of any reported characteristics.

Recommendation

Researchers should briefly report what sample characteristics they have measured and what their theory- and evidence-based reasons were for believing that these characteristics and not others might influence the IVs and/or DVs in the study. The measures should be used to analyze effects of sample choice as the following steps recommend.

V. STEP TWO: EXAMINE VARIABILITY OF SAMPLE CHARACTERISTICS

The second step in our approach (Figure 1) is to examine the variability of these sample characteristics. Variability information plays two important roles in addressing questions about participant sample choices. First, it can support prima facie judgments about the likelihood that the sample is appropriate. If the target population is relatively homogeneous (diverse) with respect to a potentially influential characteristic (e.g., nationality or IS expertise), a representative sample will be similarly homogeneous (or its diversity will be similarly distributed) with respect to this characteristic.

Second, especially when the distribution of some potentially influential characteristic differs between the target population and the sample, the variability of sample characteristics plays an important role in the testing that enables researchers to provide persuasive support for their sample choices and their judgments about the resulting limitations (see steps 3 and 4 for more detail on these tests and their uses). Too little variability in a characteristic can make it impossible to conduct meaningful tests of association between the characteristic and the variables of interest in the study. Even when variation in the characteristic is adequate for testing, the results of the test (e.g., an assurance that the characteristic has no association with the variables of interest, and thus a mismatch between sample and target population with respect to the characteristic is unproblematic) apply only to the sample range\(^5\). Researchers should therefore report on the variability and the means of potentially influential sample characteristics.

Current Practice

Variability information was limited. In the experiments we examined, only 22 percent of those that reported measures of sample characteristics also provided variability information for all these characteristics; 14 percent provided no information at all about the variability of sample characteristics, and 64 percent provided variability information for some but not all of the sample characteristics reported in the study (38% of the time, this partial information consisted only of the percentages of male and female participants in the sample). The types of variability information provided were quite diverse. We counted any of the following items as providing variability information: standard deviations, ranges, percentage distributions, and threshold values (e.g., “all participants had at least three years of work experience”). The types of variability information provided often differed across sample characteristics in and across experiments. Because variability reporting in existing studies has been infrequent, it is difficult for readers to judge how diverse the participant sample in a given study is, let alone whether the choice of diversity level is appropriate.

Recommendation

Researchers should report how variable the participant sample is with respect to all potentially influential sample characteristics. Relevant variability information includes standard deviations and marked non-normality of distribution shapes, such as bimodality or strongly skewed distributions, which might make mean values non-representative. For example, a participant group with mean software development experience of ten years may appear solidly experienced but (to take an extreme case), if one-third of the participants have around thirty years of experience and two-thirds have virtually none, the mean of ten years is misleading. Range information alone, which is often provided, has limited usefulness: for example, one or both end points of the range could be outliers that provide little information about the location of most of the observations.

---

\(^4\) A paper can report more than one experiment and can report differently on sample characteristics for each experiment. In our sample, 25 papers reported on more than one experiment. Taking this into account, we calculated this percent based on the number of experimental studies (222).

\(^5\) For example, a finding of no effect of age on the variables of interest, when ages in the sample are broadly distributed between 25 and 45, does not allow researchers to conclude that there is no effect of age on these variables in the 45-75 range.
VI. STEP 3: TEST EFFECTS OF SAMPLE CHARACTERISTICS

Step 3 is to test or otherwise assess the associations among potentially influential sample characteristics and the DVs and IVs in the study (Figure 1). The specific types and magnitudes of threats to validity posed by population and sample characteristic mismatches depend on whether and how the variance in the sample characteristic is associated with variance in the DVs and IVs in the study (see Table 4 and accompanying text below for details). These tests may, for example, allow researchers to conclude that a mismatch between sample and target population poses no threat to the validity of results. Or, conversely, the tests may indicate that the sample choice significantly limits the inferences that can be drawn from the study. Without some knowledge of the associations between sample characteristics and a study’s DVs and IVs, an informed discussion of sample choice and its consequences is not possible.
When a sample is homogeneous with respect to a certain characteristic, meaningful testing for associations between the characteristic and the DVs and IVs is not possible. In such cases, a more judgmental assessment is required. If the characteristic that is homogeneous in the sample is also homogeneous in the target population, and if mean levels are similar, then the sample is representative and sample homogeneity on this characteristic poses no threats to validity. If the target population is homogeneous at a different mean level or is more diverse, then the primary threat to validity arises from interactions between the relevant characteristic and the IVs of interest (see Table 5 below and accompanying text). In such cases, prior literature, both theoretical and empirical, can sometimes offer evidence about whether such interactions exist and thus whether the mismatch between sample and target population poses a threat to the validity of the study’s results.

**Current Practice**

Table 3 summarizes the reported testing of sample characteristics in the experimental studies we examined. Column a presents the number of experiments that measured each characteristic. Some characteristics were measured in more than one way in a single experiment. Because results of tests can differ depending on which measure was tested, we also report the total number of measures of each characteristic in square brackets in column a. The infrequency of testing and lack of significant effects that appear in Table 3; tests of experience effects on the DV were reported for 107 of the 340 measures. The infrequency of testing and lack of significant effects that appear in Table 3 shows that key tests were relatively infrequently performed. For example, 340 experience measures were collected in the studies summarized in Table 3; tests of experience effects on the DV were reported for 107 of the 340 measures. The infrequency of testing and lack of significant effects that appear in Table 3 could simply be due to the absence of variation in the characteristics. The infrequency and inconsistency of variability reporting makes it difficult to judge how often this was the case.

### Table 3: Testing of Sample Characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>(a) Number of experiments that reported a measure of the characteristic (# of measures)</th>
<th>(b) Mean difference in characteristics across experimental treatments (# tests significant)</th>
<th>(c) Additive effect of characteristics on dependent variable (# tests significant)</th>
<th>(d) Interaction of characteristics with independent variables (# tests significant)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>147 (147)</td>
<td>43 (0)</td>
<td>32 (4)</td>
<td>5 (2)</td>
</tr>
<tr>
<td>Age</td>
<td>137 (137)</td>
<td>43 (1)</td>
<td>17 (2)</td>
<td>3 (0)</td>
</tr>
<tr>
<td>Experience</td>
<td>130 (340)</td>
<td>147 (2)</td>
<td>98 (20)</td>
<td>9 (5)</td>
</tr>
<tr>
<td>Class or grade level</td>
<td>34 (35)</td>
<td>6 (0)</td>
<td>6 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Education level</td>
<td>16 (16)</td>
<td>8 (0)</td>
<td>5 (1)</td>
<td>1 (0)</td>
</tr>
<tr>
<td>Personality type variables</td>
<td>13 (13)</td>
<td>6 (0)</td>
<td>7 (2)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Academic major</td>
<td>15 (15)</td>
<td>3 (0)</td>
<td>1 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Trust disposition/risk propensity</td>
<td>12 (14)</td>
<td>8 (1)</td>
<td>6 (4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Grade point average</td>
<td>10 (10)</td>
<td>5 (0)</td>
<td>2 (2)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Ethnic background</td>
<td>8 (8)</td>
<td>2 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Personal relevance of task</td>
<td>7 (8)</td>
<td>2 (0)</td>
<td>3 (2)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Income</td>
<td>6 (6)</td>
<td>4 (0)</td>
<td>2 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>English as a second language</td>
<td>5 (5)</td>
<td>2 (0)</td>
<td>1 (1)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Citizenship</td>
<td>4 (4)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Country of birth</td>
<td>2 (2)</td>
<td>0 (0)</td>
<td>1 (1)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Voice quality</td>
<td>2 (6)</td>
<td>0 (0)</td>
<td>6 (5)</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Motivation to learn</td>
<td>1 (1)</td>
<td>1 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Chronic illness</td>
<td>1 (1)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>550 (768)</strong></td>
<td><strong>280 (4)</strong></td>
<td><strong>187 (44)</strong></td>
<td><strong>23 (12)</strong></td>
</tr>
</tbody>
</table>

6 “Number of measures” represents the total number of measurements, not the number of types of measures. Thus, if experience is measured as “months of work experience” in thirty studies, this counts as thirty measures.
However, it is unlikely that absence of variation accounts for most of the low level of reported testing, which we can see by examining sample characteristics that are likely to vary in experimental samples (i.e., gender and, for student samples, academic major). Because 147 experiments in our sample reported gender information, 441 tests could, in principle, have been performed (147 experiments x the 3 tests in columns b–d of Table 3). However, only 80 tests (18% of the possible number) were reported. Similarly, 15 experiments recorded participants’ academic major, which results in 45 possible further tests, but the studies reported only four such tests (9%).

Insofar as tests on sample characteristics were reported at all, the most common test was for mean difference in the characteristics across experimental treatments (i.e., for correlation between the sample characteristic and levels of the IV). Thirty-six percent (280) of the 768 instances of measured characteristics were tested for correlation with the IVs in the study. When researchers randomly assigned participants to the treatments, no such correlations would be expected, although differences can occur with a sufficiently bad draw from the random distribution. For example, in Sia, Tan, and Wei (2002), mean age differs significantly between groups of participants randomly assigned to face-to-face and dispersed conditions of computer-mediated communication. Age is then used as a control variable (covariate) in this study’s hypothesis tests to provide assurance that the apparent effect of different settings on group processes is not due instead to differences in age. Over half the studies in our sample provided no assurance on this point.

Sample characteristics are more likely to be correlated with measured IVs than with manipulated, randomly assigned IVs. For example, when experience of a particular type is the IV of interest, it may well be correlated with age, amount and type of education, and various personality or attitude characteristics that influence individuals’ dispositions to take certain jobs and acquire certain experience. When such correlations exist, they can raise questions about the validity of the tests of IV effects on the DV. Although 36 of the 184 papers in our sample included measured IVs, only two reported testing for correlations between the sample characteristics and the IVs.

Sample characteristics that influence the DV can provide alternative explanations for DV variance, which competes with the explanation provided by the IVs in which the researcher is interested. Or they can limit the results’ range of generalizability (see Section 7 for details). These effects of sample characteristic on DVs can be either additive (independent of IV effects) or interactive. Ideally, researchers should test for both. Among the 768 measures of sample characteristics we identified, we found only 187 (24%) instances of tests for additive effects on the DVs. Testing for interactive effects of sample characteristics was even rarer, being reported for only 23 (3%) of the 768 measured characteristics.

**Recommendations**

When there is sufficient variance in sample characteristics to make such tests meaningful, researchers should report tests of sample characteristics’ correlations with IV(s) and their additive and interactive effects on the DV(s). This information about the relations between sample characteristics and other IS-related variables can be valuable not only for analyzing threats to the validity of a study’s results, but also for guiding future participant choices and assisting researchers who want to build and test theory related to these characteristics. When a potentially influential characteristic is homogeneous within the sample, researchers should report their reasons for believing that the characteristic either is similarly homogeneous in the population or does not interact with the IVs in the study. Such tests and assessments were relatively infrequent in the studies we examined, and the absence of variability information often made it difficult to determine when each approach would have been appropriate.

**VII. STEP 4: IDENTIFY AND REPORT VALIDITY THREATS**

Step 4 is to identify and report validity threats resulting from the distribution of potentially influential characteristics in the sample (Figure 1). We focus on three of the four validity types that Shadish et al. (2002) define:

1. **Statistical conclusion validity:** do the DVs and IVs actually covary (or not) when the statistical tests in the study indicate that they do (or do not)?

2. **Internal validity:** is the observed covariance causal?

---

7 We do not list all the substantive sources of threats to validity that appear in Shadish et al. (2002), such as “selection”, “history”, etc. Instead, we provide a compact, structured framework for thinking about these substantive issues in terms of statistical inference problems.

8 This view of internal validity assumes that the hypotheses being tested are causal in intent (e.g., computer-mediated communication causes different behavior than face-to-face communication), rather than hypotheses about parameter values (e.g., the mean return on investment in IT is greater than 10%).
3. External validity (generalizability): do the study's conclusions generalize beyond the sample and experimental setting employed?

The fourth validity type that Shadish et al. (2002) identify is construct validity, which is whether the measured or manipulated variables capture the theoretical constructs of interest in the study. We do not include construct validity here because characteristics of the individuals included in the sample typically do not influence the construct validity of the variables of interest in the study. For example, if the DV is a questionnaire measure of trust in an information system or intention to use it, there is typically little reason to believe that the questionnaire will be a good measure of the construct for men but not for women, or will be a good measure for individuals in their twenties but not individuals in their forties.

Table 4 defines and presents examples of five sample characteristic types with differing statistical properties. We first describe these five types and then show how researchers can use this typology to identify the specific threats to validity that are posed—and often not posed—by imperfectly representative samples.

<table>
<thead>
<tr>
<th>Sample characteristic</th>
<th>Relation to IV and DV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1: uncorrelated</td>
<td>Uncorrelated with any IVs and does not influence DV</td>
</tr>
<tr>
<td>Type 2: IV-correlated only</td>
<td>Correlated with one or more IVs but does not influence DV</td>
</tr>
<tr>
<td>Type 3: DV-correlated only</td>
<td>Influences DV but is not correlated with IVs and does not interact with any IVs to influence DV</td>
</tr>
<tr>
<td>Type 4: IV- and DV-correlated</td>
<td>Influences DV and is correlated with one or more IVs, but does not interact with any IVs to influence DV</td>
</tr>
<tr>
<td>Type 5: interacting</td>
<td>Interacts with an IV to influence DV, and may or may not be correlated with IVs</td>
</tr>
</tbody>
</table>

Types 1 and 2 are characteristics that are potentially associated with the DVs and/or IVs of the study. That is, researchers believe that the characteristics might be associated with the DVs or IVs a priori. But testing reveals that, even with reasonable variation in the characteristic, the characteristics are not associated with the DV(s)—and in type 1, with the IVs. For example, in their investigation of e-commerce trust, Kim and Benbasat (2006) found that a sample characteristic, online shopping frequency, was uncorrelated with trust-assuring argument displays (the IV), and had no statistically significant association with trust beliefs (the DV). It was reasonable ex ante to suppose that such correlations might exist and hence the researchers tested for them. But the tests demonstrated no relation between the sample characteristic and the variables of interest in the study.

Type 2 sample characteristics are correlated with one or more IVs but do not influence the DV. That is, controlling for the IVs, the characteristic has no significant incremental association with the DV. For example, in Sia et al. (2002), age was a type 2 sample characteristic. Age was significantly higher in one of the experiment's computer-mediated communication treatments than in the other (thus it was correlated with the computer-mediated communication IV). However, age had no significant incremental association with the DVs (choice shift and preference change).

Type 3 sample characteristics influence one or more DVs, but do not correlate with any IVs or interact with them to influence the DV (i.e., their effect on the DV is additive). As an example of a type 3 sample characteristic, Piccoli, Ahmad, & Ives (2001) found that gender had a significant influence on the DVs (performance, satisfaction, and self-efficacy) and did not differ between learning environment treatments (virtual vs. traditional). While the authors did not report the results of any interaction tests, we assume for the convenience of this example that there are no interactions between gender and the learning environments.

Type 4 sample characteristics are identical to type 3 sample characteristics except that they are correlated with one or more IVs. Mennecke, Crossland, and Killingsworth (2000) provide an example of a type 4 sample characteristic. In their study, the sample characteristic “task interest” significantly influenced the DV (time spent on the task) and was also correlated with one of the IVs, expertise.

---

9 However, errors in measuring the sample characteristics (a problem analogous to construct validity issues for IVs and DVs) can threaten the validity of the study's inferences via threats to statistical conclusion validity or internal validity. These threats are described in the following sections.
Type 5 sample characteristics interact with an IV to influence the DV, as illustrated in Allen and March (2006). They found that a sample characteristic, comfort level in writing queries, interacted with the treatment (ontological foundation) to significantly influence the DV (prediction of accuracy)\(^\text{10}\).

Researchers can use the five sample characteristic types shown in Table 4 to identify different threats to validity arising from potentially influential sample characteristics\(^\text{11}\). Table 5 presents the potential threats to the three validity types that can occur for each sample characteristic type. The specific threats that occur depend on whether the characteristic is relatively diverse or homogeneous within the sample and, if diverse, whether the characteristic is included in the empirical model used to test hypotheses\(^\text{12}\).

<table>
<thead>
<tr>
<th>Sample characteristic types for which threat can occur (see Table 4)</th>
<th>Experiment participant and empirical model choices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diverse sample characteristics not measured and not included in model</td>
</tr>
<tr>
<td>Type 1</td>
<td>Statistical conclusion validity threats (reduced degrees of freedom)</td>
</tr>
<tr>
<td>Type 2</td>
<td>Statistical conclusion validity threats (reduced degrees of freedom, multicollinearity)</td>
</tr>
<tr>
<td>Type 3</td>
<td>Statistical conclusion validity threats (unexplained variability in Y)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 4</td>
<td>Statistical conclusion validity threats (unexplained variability in Y)</td>
</tr>
<tr>
<td></td>
<td>Internal validity threats (correlated omitted variable)</td>
</tr>
<tr>
<td>Type 5</td>
<td>Statistical conclusion validity threats (unexplained variability in Y)</td>
</tr>
<tr>
<td></td>
<td>Internal validity threats (correlated omitted variable, aggregation error)</td>
</tr>
<tr>
<td></td>
<td>External validity threats</td>
</tr>
</tbody>
</table>

\(^{10}\) Type 5 sample characteristics may or may not be correlated with IVs. If they are correlated, then the correlation raises issues similar to those for Type 4 sample characteristics. Therefore, in discussing type 5 sample characteristics, we focus only on the interaction implications.

\(^{11}\) Each threat to validity is represented independently in our discussion. For example, in presenting threats to external validity (generalizability), we assume that a statistically and internally valid inference has been drawn. The focus is then on whether the inference is also valid for settings outside the laboratory and individuals other than the participants actually used in the experiment.

\(^{12}\) When a characteristic is homogeneous in the sample, its lack of variance will insure that it has no significance in the model. Hence, including it will be uninformative.
Diverse Samples: Sample Characteristics Not Measured and Not Modeled

It is sometimes difficult to be certain about what the relevant sample characteristics and their statistical properties are in a population of interest, and good measures and models for relevant characteristics are not always available. Perhaps because of these difficulties, sample characteristics are often omitted from empirical models used in hypothesis testing, which Table 3 indicates. With diverse samples, this can pose threats to all three of the types of validity we consider here.

First, if the omitted characteristics influence the DV (characteristic Types 3, 4, or 5), they will create unexplained variability in the DV (a large error term in the model). As such, they will weaken the power of statistical tests and create threats to statistical conclusion validity. Larger samples and/or stronger manipulations that produce larger mean effects are straightforward ways of dealing with this threat. Other threats resulting from unmeasured diversity are not so easily mitigated, however.

Second, if the omitted characteristics are types 4 or 5, then they are correlated omitted variables, which can pose important threats to internal validity. Omitting a variable that is positively correlated with an IV inflates the estimated coefficient on the IV in the empirical model, which potentially results in a significant coefficient even when the IV has no causal influence on the DV. Conversely, omitting a variable that is negatively correlated with an IV reduces the estimated coefficient on the IV, which potentially results in a non-significant coefficient even when the IV actually has a significant influence on the DV (see MacKinnon, Krull, and Lockwood (2000) for a detailed discussion of the effect of correlated variables).

The third threat to validity from diverse samples with omitted sample characteristics is a threat to generalizability (external validity) that occurs with type 5 (interacting) characteristics. If sample characteristics are not measured, it is difficult to judge how representative a diverse sample is. The proportion of individuals with high and low values on a particular characteristic may very well differ between the participant sample and the target population. The numerical example in Table 6 illustrates how this can lead to generalizability problems for type 5 (but not other) characteristics.

### Table 6: Example of Additive Effects versus Interaction Effects

<table>
<thead>
<tr>
<th>Types 3 and 4 sample characteristics: additive effects on Y</th>
<th>Type 5 sample characteristic: interaction effect on Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecasting performance (Y)</td>
<td>Low problem-solving ability</td>
</tr>
<tr>
<td>Without DSS</td>
<td>40</td>
</tr>
<tr>
<td>With DSS</td>
<td>60</td>
</tr>
<tr>
<td>Means</td>
<td>50</td>
</tr>
<tr>
<td>Cell entries are forecasting performance on a 0–100 scale</td>
<td></td>
</tr>
</tbody>
</table>

In this example, the sample characteristics (problem-solving ability and statistical knowledge) are characterized for simplicity as either high or low, and both have significant effects on the DV, forecasting performance. In the additive example (a type 3 or 4 characteristic), mean forecasting performance is higher by twenty points with a decision support system (DSS) than without it—and this is true for both low-ability and high-ability individuals. In the interaction example (a type 5 characteristic), in contrast, the mean effect of the DSS on performance is not the same for individuals with high and low values of the characteristic. Those with high statistical knowledge forecast more accurately when they use a DSS than when they do not, but those with low statistical knowledge are more accurate without the DSS, perhaps because the DSS requires statistical knowledge for effective use and confuses individuals with low knowledge.

If the type 5 characteristic, statistical knowledge, is not included in the model, and if roughly equal numbers of high- and low-knowledge individuals are in the sample, then mean forecasting performance will appear identical with and without the DSS, which the marginal means on the right-hand side of Table 6 show. Not only is this result an incorrect inference about the real (non-zero) effect of the DSS (an internal validity problem due to the omitted interaction variable), but it also has limited external validity. The conclusion that the DSS has no mean effect on forecasting performance will not generalize to any population that is not a 50-50 mix of low- and high-knowledge

---

13 Even if type 5 sample characteristics are not themselves correlated with an IV, an interaction between a characteristic and an IV will be correlated with the IV. Thus, when the characteristic is not measured and included (as an interaction term) in the empirical model, the interaction will be a correlated omitted variable.
individuals. Moreover, if there is a clear separation between individuals who have enough knowledge or perform well with the DSS and those who do not, then the mean effect of DSS use in a mixed-experience group does not generalize to any individual because no individual has 50 percent high knowledge and 50% percent low knowledge. Thus, the null main effects in this example are aggregation errors and generalize neither to different samples or populations, nor to any specific individuals (see Lynch, 1999 for a discussion of aggregation errors).

Diverse Samples: Sample Characteristics Measured and Modeled

The threats to validity that are posed by omitting potentially relevant sample characteristics can, in principle, be obviated by measuring these characteristics and including them in the empirical models used for testing hypotheses. However, especially when sample-characteristic measures are imperfect and/or the relations of the characteristics with IVs and DVs are uncertain, including the characteristics in empirical models can create other threats to statistical conclusion validity and internal validity.

First, if a number of additional sample characteristics with little explanatory value (types 1 and 2) are included in a model “just in case”, then the model can lose power because the additional variance explained by the sample characteristics is not sufficient to compensate for losing degrees of freedom. Although it is straightforward to solve this problem by re-estimating the model without the uninformative characteristics, other threats are not so easily mitigated.

A second threat, multicollinearity, arises when characteristics are correlated with one or more IVs (type 2, 4, and 5). Multicollinearity inflates standard errors of the IV coefficient estimates, which results in threats to statistical conclusion validity due to imprecise coefficient estimates and low-power hypothesis tests. Widely used rules-of-thumb for identifying multicollinearity problems tend to understate these problems. For example, it is common to regard variance inflation factors (VIFs) greater than 5 or 10 as indicators of multicollinearity problems. But lower VIFs can still require a doubling or tripling of sample size, relative to a sample without correlations among the IVs and sample characteristics, to maintain adequate power (see Hsieh, Bloch, and Larsen 2003 for further information on VIFs and sample sizes). Note that, if a sample characteristic is highly correlated with the IV, it can be difficult for researchers to determine whether it has incremental explanatory power for the DV (a type 4 or 5 characteristic rather than a type 2 characteristic) using only sample data because the correlation will make it difficult to disentangle effects of the sample characteristic and the IV. In such cases, theory that helps to judge the likelihood of a causal relation between the sample characteristic and the DV, and empirical evidence from other studies with different correlation structures in the sample, can provide a basis for classifying the characteristic as type 2 or a type 4 or 5.

Third, measurement and specification errors related to the sample characteristics can pose threats to valid inference when researchers use diverse samples and include the characteristics in empirical models including characteristic types 3, 4, and 5. For example, measures of individuals’ experience and knowledge often capture the underlying constructs with some error, especially when the construct researchers want to control for is task-relevant knowledge and the measure is years of work experience. This error can create both inconsistency and bias in the estimated coefficients, which results in threats to both statistical conclusion validity and internal validity (Greene, 2000; Wooldridge, 2006)\textsuperscript{14}.

Errors in specifying the functional form of the relation between a sample characteristic (types 3, 4, and 5) and the DV can also bias the coefficients on the IVs. For example, suppose that there are diminishing returns to experience (a curvilinear relation between task experience and task performance that can be represented by an experience-squared term in the model). Suppose further that the relation between experience and performance is modeled as linear in the empirical analysis, with the quadratic term omitted from the model. If experience (and thus the omitted experience-squared term) is correlated with an IV in the empirical model, then the experience-squared term is a correlated omitted variable, with the potential to bias the estimated coefficient on the IV. Moreover, when a sample characteristic has a curvilinear effect on the DV and the quadratic term is omitted from the model, tests for interaction effects can be distorted: they can show no effect when an interaction actually exists or show a different form of interaction than actually exists (Ganzach, 1997). Because interaction effects are important both for understanding how an IV affects the DV in the sample and for identifying limits to the generalizability of the effect, this specification problem can be significant.

Thus, using diverse, representative samples and controlling for sample-characteristic effects in empirical models is not always an effective strategy for avoiding significant threats to valid inference. In particular, multicollinearity and

\textsuperscript{14} The existence and nature of these threats depends on the structure of the correlations among the measurement error and the IVs. Even when this correlation structure is known, “the sizes and even the directions of the biases (in coefficient estimates) are not easily derived” (Wooldridge, 2006, p. 320).
sample-characteristic measurement or model-specification errors can result in invalid inferences about the IV–DV relations a study investigates.

**Homogeneous Samples**

Participant samples that are relatively homogeneous with respect to relevant characteristics\(^{15}\) avoid the threats to valid inference that arise with more diverse samples. Characteristics that vary little in the sample will not create unexplained variance in the DV, nor will they create variance incorrectly attributed to an IV if they are omitted from the model. Because the sample characteristics do not need to be included in the model, characteristic measurement and model-specification errors do not threaten the validity of hypothesis tests.

The usual concern about homogeneous samples is external validity, but, as Table 4 indicates, only one type of sample characteristic—type 5, which interacts with an IV—actually creates external validity threats. External validity concerns do not arise with non-interacting characteristics, even when they have significant effects on the DVs, and the sample and target population differ significantly with respect to the characteristic. In the hypothetical example of additive effects in Table 6, individuals with low problem-solving ability do not forecast as well as individuals with high forecasting ability. But the effect of DSS use on forecasting performance for low-ability individuals (a 20-point improvement) generalizes to high-ability individuals and vice versa. Thus, the failure to match target population and sample on characteristics with additive effects only (types 3 and 4) has no effect on external validity\(^ {16}\). An experiment using only low-ability, only high-ability, or any mix of low- and high-ability individuals will show a mean DSS effect of 20 points on forecasting performance. In consequence, when effects are additive, the researcher defining a target population does not need to specify its ability level or find a participant sample that has the same ability level or mix as the target population. Only the type 5 (interacting) characteristic, illustrated on the right-hand side of Table 5, poses a threat to external validity when sample and target population are not matched with respect to the characteristic.

**Current Practice and Recommendations**

Because of the close relation between identifying and reporting threats to validity (step 4) and reporting trade-offs based on the importance of the threat(s) (step 5), we provide current practice and recommendations for both steps at the end of Section 8.

**VIII. STEP 5: ANALYZING AND REPORTING SAMPLE-CHOICE TRADEOFFS**

Sample choice, like other elements of research design, often cannot minimize all threats to validity simultaneously at a reasonable cost. Even well-conducted studies will often fail to eliminate some threats to validity. When researchers have identified the threats to validity that their study has not eliminated, their task is then to explain how their sample choice represents an appropriate tradeoff; that is, how they are accepting some smaller threats to reduce other larger ones.

Researchers who consider the statistical characteristics summarized in Table 4 and the resulting specific threats to validity (Table 5) can help themselves judge which threats are likely to be large and which are not. For example, the magnitude of threats to statistical conclusion validity resulting from power limitations will be a relatively large threat when specialized study requirements keep sample sizes small, but not when large samples can overcome the power problems. The magnitude of threats to external validity resulting from unrepresentative samples will be large when existing literature indicates a likelihood of IV x sample characteristic interactions (or this likelihood is altogether unknown), but not when existing literature provides evidence against interactions.

**Current Practice**

Many studies report on participant choice and its consequences with qualitative comments in their methods or conclusion sections. To some extent these qualitative comments address validity threats, but often not in ways that can clearly be matched to each of the specific threats identified in Table 5 or to the tradeoffs among them. In this section, we identify the five types of qualitative comments that we found most frequently reported in the studies we analyzed. For each, we specify common limitations of the comments, which, in many cases, could be readily overcome, which makes these comments more informative and persuasive.

---

\(^{15}\) That is, sample characteristics that are statistically associated with an IV and/or DV in the population or in a diverse sample.

\(^{16}\) Recall also that failure to match sample and target population on type 2 characteristics, which have no effect on the DV and are correlated with an IV, will have no effect on external validity.
Comment 1. Homogeneity: IS researchers occasionally cite the value of a homogeneous sample in reducing extraneous variation in behavior. This is an important point, consistent with the value of homogeneity in reducing threats to validity as represented in Table 5.

Limitations: This type of qualitative comment will be more persuasive when it is supported by two kinds of evidence:

- variability information on relevant sample characteristics that confirms the homogeneity, and
- evidence (so far as it is available, either from other research or from such variability of the sample characteristic as exists in the study) that the IVs in the study do not interact with the sample characteristics over the range of the characteristic observed in the target population.

Comment 2. Sample size: For studies that require large sample sizes, the participant choice is sometimes justified by the observation that large numbers were easy to obtain with a particular participant group. Sufficiently large sample sizes can reduce the threats to statistical conclusion validity that arise from the various sources of reduced statistical power that are summarized in Table 5.

Limitations: If sample-size considerations are the only defense provided for participant choice—as they sometimes are—they are insufficient. A large sample size does not help with the internal validity (biased hypothesis tests) or external validity (generalizability) concerns summarized in Table 5. Hence, authors should provide some assurance about these threats to persuade readers that the study’s participant choice represents a favorable tradeoff between statistical power and other validity concerns.

Comment 3. Prior valid use of a participant group: IS researchers sometimes justify the choice of a particular participant group based on the fact that the group has been used in prior high-quality IS research, and results have tended to generalize to diverse populations.

Limitations: This justification is not very informative unless it is specific. For example, suppose that the concern about participant choice in a particular study is whether the results are generalizable from a low-experience participant sample to a population that includes higher-experience individuals. The concern is whether experience interacts with the IVs of interest in the study. The fact that experience does not interact with different IVs used in prior literature is not very informative on this point. To provide a defense for the choice of participants, the prior research that is cited needs to share IVs with the study in question, and the prior research needs to indicate that experience does not interact with these IVs.

Comment 4. Participants are real users: IS researchers sometimes present their participant choice as appropriate because the study is about (for example) online consumers or decision support system users, and the participants are in fact online consumers or decision support system users.

Limitations: Results from one group of “real” users do not necessarily generalize to other groups of “real” users. For example, Ko and Dennis (2011) find that how much knowledge workers in an organization benefit from a knowledge management system depends on the workers’ specific prior job experience and when the research data are gathered (shortly after the introduction of the system or later). If the authors had been able to gather data only at one point in time, or only from knowledge workers with a narrow range of experience, then the fact that their participants were “real” knowledge workers would have been no guarantee of generalizability. All the validity tradeoff concerns summarized in Table 5 apply equally to all samples, and observing that the participants are “real” is not a substitute for analyzing these tradeoffs.

Comment 5. Recommending replications: Some IS experimental studies acknowledge that their participant samples are limited and therefore their results should be replicated with different samples.

Limitations: Although replications can be valuable, general recommendations for replication can be uninformative, in that they can leave the impression that the key to generalizability is simply testing over and over again with differing samples. Building theory- and evidence-based arguments identifying the sample characteristics that are likely to interact with the variables of interest and focusing on testing these interactions is likely to be a more fruitful strategy. Hence, recommendations for replications are likely to be more meaningful when they are more clearly focused on potential interaction issues.
Recommendation

A good research design (including sample, measurement, and modeling choices) is one that avoids large threats, though possibly at the expense of incurring smaller threats. These tradeoffs should be considered when researchers embark on a study, and we recommend that IS researchers explicitly discuss the tradeoffs between specific validity threats that result from the participant choices made in their studies to persuade readers that their sample choice is appropriate. The smaller threats should be appropriately identified as limitations.

For example, if theory and previous research (using diverse samples) demonstrate that there are no interactions (type 5) between a certain sample characteristic and the independent variables in the study, then researchers can select a sample that is homogenous with respect to that characteristic (and measure the characteristic to document the homogeneity) because this avoids the other potential threats to validity. Conversely, if theory supports the existence of an interaction, or interaction effects have been demonstrated, then researchers should select a representative sample with regard to that characteristic and carefully measure it because this is the only way to identify the limits to generalizability that are created by interacting (type 5) characteristics. When it is uncertain ex ante whether a characteristic interacts with a study’s independent variables, then researchers need to make—and explain—their judgments about the likelihood of interaction, the importance of broad generalizability in their study, and the likely magnitude of the threats to validity that can result from using a diverse sample. For example, if a characteristic can be reliably measured and modeled (and this is documented in the study), then the threats to validity that arise from using a diverse sample are lower, and a diverse sample can be attractive even when the likelihood of interaction is only moderate. If well-validated measures and models are not available and the reasons to expect interaction are weaker, then the validity tradeoffs can favor a homogeneous sample.

V. CONCLUSION

Based on our proposed framework, we analyze experimental studies appearing in four major IS journals. We identify and analyze their participants, target populations, and tasks (Table 2), and show that there is not always a prima facie match between samples and target populations. We provide a prescriptive, step-by-step approach for testing and reporting on potential threats to validity resulting from participant choice (Figure 1). We show that current practice dealing with sample characteristic variation in IS experiments has some features that are valuable and should be maintained. For example, researchers sometimes collect sample characteristic data and perform tests (e.g., for equality of means across experimental treatments) that help to allay concerns about the validity of study inferences. We also show where current practice in IS research could be improved.

As our analysis illustrates, the questions that arise about sample characteristic influences are often both methodological questions about how best to conduct and analyze experiments and substantive questions about the identity and form of significant influences on IT-related behavior. Future experiments can test and refine theory in this area, and more refined theory can be used to support more informative experimental tests in the future. Efforts toward “realism” in participant choice by simply using more diverse or more experienced participants are likely to be less helpful than building a better theory-based understanding of the relations between sample characteristics and IV/DVs in IS research. As Lynch (1999, p. 368) observes:

The only path to understanding the generality of one’s findings is to have a theory that specifies moderator (interacting) variables and boundary conditions and specifies what variables should not moderate the findings reported, and to test for the asserted pattern of interactions. If one’s theory is impoverished, no degree of adherence to methodological prescriptions will help ‘ensure’ external validity.

REFERENCES

Author’s note: The complete list of all papers in our sample is in Appendix A.

Editor’s Note: The following reference list contains hyperlinks to World Wide Web pages. Readers who have the ability to access the Web directly from their word processor or are reading the paper on the Web, can gain direct access to these linked references. Readers are warned, however, that:

1. These links existed as of the date of publication but are not guaranteed to be working thereafter.
2. The contents of Web pages may change over time. Where version information is provided in the References, different versions may not contain the information or the conclusions referenced.
3. The author(s) of the Web pages, not AIS, is (are) responsible for the accuracy of their content.
4. The author(s) of this article, not AIS, is (are) responsible for the accuracy of the URL and version information.


**APPENDIX A: EXPERIMENTAL STUDIES**


ABOUT THE AUTHORS

Nancy Lankton is an Associate Professor in the Division of Accountancy and Legal Environment at Marshall University. She teaches accounting information systems and information systems auditing. Nancy’s main research interests include trust’s impacts on individual and organizational use of information technology, privacy behaviors related to social networking websites, and decision making in organizations. She has published in journals such as *Contemporary Accounting Research, IEEE Transactions on Engineering Management, Journal of Management Information Systems, and the Journal of the American Medical Informatics Association*. She is a member of the Association for Information Systems, and the Information Systems Auditor and Control Association, and serves as an Associate Editor for Communications of the Association for Information Systems.

Joan Luft is Eli Broad Professor (emerita) of Accounting at Michigan State University. Her research, which focuses on experimental studies in accounting, has appeared in *The Accounting Review, Journal of Accounting and Economics, Accounting, Organizations and Society, Contemporary Accounting Research*, and other journals. She has been Editor or Associate Editor of *The Accounting Review, Journal of Management Accounting Research, and Accounting Horizons*.

Copyright © 2014 by the Association for Information Systems. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than the Association for Information Systems must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or fee. Request permission to publish from: AIS Administrative Office, P.O. Box 2712 Atlanta, GA, 30301-2712, Attn: Reprints; or via e-mail from ais@aisnet.org.
**EDITOR-IN-CHIEF**
Matti Rossi  
Aalto University

**AIS PUBLICATIONS COMMITTEE**

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
</table>
| Virpi Tuunainen       | Matti Rossi  
Aalto University                      |
| Vice President Publications  
Aalto University | Suprateek Sarker  
Editor, JAIS  
University of Virginia |
| Robert Zmud           | Phillip Ein-Dor  
AIS Region 2 Representative  
Tel-Aviv University |
| AIS Region 1 Representative  
University of Oklahoma | Bernard Tan  
AIS Region 3 Representative  
National University of Singapore |

**CAIS ADVISORY BOARD**

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
</table>
| Gordon Davis          | Ken Kraemer  
University of California at Irvine |
| Jay Nunamaker         | Henk Sol  
University of Groningen              |
| M. Lynne Markus       | Ralph Sprague  
University of Hawaii                 |
| Richard Mason         | Hugh J. Watson  
University of Georgia                 |

**CAIS SENIOR EDITORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
</table>
| Steve Alter           | Michel Avital  
University of San Francisco Copenhagen Business School |

**CAIS EDITORIAL BOARD**

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
</table>
| Monica Adya           | Dinesh Batra  
Florida International University |
| Marquette University                           |
| Tina Blegind Jensen  | Copenhagen Business School           |
| Indranil Bose         | Indian Institute of Management Calcutta |
| Thomas Case           | Tom Eikebrokk  
University of Agder                  |
| University of Hamburg | George Southern University            |
| Andrew Gemino         | Mary Granger  
George Washington University          |
| Simon Fraser University | Matt Germonprez  
University of Nebraska at Omaha     |
| Shuk Ying (Susanna) Ho | Jonny Holmström  
Umeå University                    |
| Australian National University | Tom Horan  
Claremont Graduate University       |
| K.D. Joshi            | Michiel Kalika  
University of Paris Dauphine         |
| Washington State University | Karlheinz Kautz  
Copenhagen Business School           |
| Nelson King           | Hope Koch  
Baylor University                   |
| American University of Beirut | Nancy Lankton  
Marshall University                 |
| Paul Benjamin Lowry   | Fred Niederman  
St. Louis University                 |
| City University of Hong Kong | Shan Ling Pan  
National University of Singapore    |
| Katia Passerini New Jersey Institute of Technology | Jan Recker  
Queensland University of Technology |
| Saonee Sarker         | Raj Sharman  
State University of New York at Buffalo |
| Washington State University | Thompson Teo  
National University of Singapore    |
| Arvind Tripathi       | Frank Ulbrich  
Newcastle Business School           |
| University of Auckland Business School | Chelley Vician  
University of St. Thomas            |
| Fons Wijnhoven        | Vance Wilson  
Worcester Polytechnic Institute       |
| University of Twente                           |
| Yajiong Xue           | Ping Zhang  
East Carolina University             |
| Padmal Vitharana      | Syracuse University                  |
| Jim Stein             | Syracuse University                  |
| Jeremy Rees           | Purdue University                    |
| Steven Short           | Miami University                     |
| Jacky Rees             | Julie Kendall  
Rutgers University                  |
| Nelson King            | Claudi Loebbeck  
University of Cologne                |
| New York University of Technology | St. Louis University   |
| Virginia Commonwealth University | East Carolina University |

**DEPARTMENTS**

<table>
<thead>
<tr>
<th>Department</th>
<th>Editor</th>
</tr>
</thead>
</table>
| Debate                                      | History of Information Systems  
Editor: Ping Zhang |
| Information Systems and Healthcare          | Information Technology and Systems  
Editors: Dinesh Batra and Andrew Gemino |
| Administrative                              | Copyediting by  
Adam LeBroq, AIS Copyeditor |

**ADMINISTRATIVE**

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
</table>
| James P. Tinsley      | Meri Kuikka  
CAIS Managing Editor  
Aalto University |
| AIS Executive Director |                                                                 |

**ISSN:** 1529-3181