Methodological Issues in Experimental IS Research: Experiences and Recommendations

By: Sirkka L. Jarvenpaa
Gary W. Dickson
Gerardine DeSanctis
Department of Management
Sciences
University of Minnesota
Minneapolis, Minnesota

Abstract

Within the last ten years the use of experimental methodology in information systems (IS) research has substantially increased. However, despite our experience with experimentation, studies continue to suffer from methodological problems. These problems have led to an accumulation of conflicting results in several areas of IS research. Moreover, future research studies will keep producing contradictory results unless researchers begin to answer questions of task and measurement validity before reporting their experimental findings. This article discusses common methodological problems in experimental IS studies and, through a description of a series of graphics experiments at the University of Minnesota, illustrates the particularly acute problem of low internal validity. Suggestions are offered to experimental IS researchers on how some of these common problems can be alleviated or even avoided, particularly in studies on the use of managerial graphics.

Keywords: MIS research, research methodology, computer graphics

ACM Categories: A.0, H.1.0, H.4.2, I.3.0.

Introduction

Experimental research has become one of the most popular forms of information systems (IS) research. The popularity and viability of an experimental approach was undoubtedly given impetus by the “Minnesota Experiments” [14] which were performed throughout the early 1970s. The researchers at Minnesota found the experimental methodology to be useful in investigating cause and effect relationships between attributes of the decision maker, the nature of the decision environment, the characteristics of an information system, and decision performance. Since the mid-seventies an increasing number of researchers have adopted the experimental approach in accumulating knowledge on the design, development, and management of information systems. Despite this surge of interest in experimentation during the past 10 years, the research suffers from methodological weaknesses, particularly from the problems of reliability and internal validity. Reliability refers to errors in measurement, whereas internal validity deals with improper manipulation of experimental treatments. The existence of these problems has led to conflicts in reported results [e.g., see 13, 22, 25].

To illustrate the situation consider one area of IS research that is struggling due to conflicting results from experimental studies — the study of information presentation formats. Several researchers have generated results that suggest that graphs are no better than tables in presenting information [19, 26, 27, 34, 36, 51]. On the other hand, several experiments have provided evidence that graphics are not only preferred by managers but also lead to better decision performance in some situations [2, 18, 38, 53, 54]. As a further illustration, three studies using interpretation accuracy as a dependent variable were in disagreement. Feliciano, et al. [17] concluded that graphs are easier to interpret than tables. Lusk and Kersnick [28] found the reverse, and later Tullis [48] demonstrated that there was no difference in interpretation accuracy between graphs and tables.

What is happening here? It is the authors' contention that there are a variety of factors leading to this confusing situation. One important prob-
Problem is that a variety of tasks are used across experimental studies, and it very likely is inappropriate to compare results in one task environment with those in another. A second problem is that the quality of information presentation across different experiments varies greatly. One study, for example, may use high quality graphics whereas another employs graphics of questionable quality. Differences in subject characteristics, experimental conditions, settings, and design likewise make conclusions about the performance of various presentation media difficult to draw. However, the main reason for inconsistent results, as argued in this article, is that the results being reported are based upon experiments in which internal validity is questionable. In other words, the experimenter has not carefully insured that the outcomes being observed are truly a function of experimental manipulation of the independent variables, or that the experiments employed adequately and accurately measure what is intended. Seldom do researchers present the reader with any discussion of how the task and measurement devices were tested.

This article is directed toward information systems researchers in general, but may be of particular interest to those contemplating experimental research in the study of managerial graphics. We will use the latter area, one in which we are initiating a program of research, to illustrate our position. The article will use three experiments as the basis for arguments concerning the methodological issues, particularly those involving internal validity. Under normal circumstances the reader might expect to see the results of such experiments as the basis for anywhere from one to three papers. We are not choosing this approach since we question the validity of the results we would be reporting, and it is our feeling that this would simply add to the confusing situation now present in the literature. We begin by discussing the common problems of experimental validity, highlighting the problems of task design and measurement. We then describe the process we are employing to address issues of internal validity in experimental studies of managerial graphics. The article concludes by providing some cautions and guidelines for experimental IS researchers.

Problems in IS Experimental Research

Taylor and Benbasat [47] describe a series of methodological problems plaguing one area of experimentally based IS research — the role of cognitive style in decision making. These problems, which generally apply to much IS experimental research, are: (1) lack of theories for guiding research efforts, (2) proliferation of measuring instruments, many of which lack reliability and validity, and (3) inappropriate research designs. In addition to these we have another problem which involves (4) inconsistency in the task that serves as the basis for an experiment. Since the resolution of all four of these problems is a prerequisite for advances in scientific endeavor, we will explore the nature of each problem and introduce some potential remedies. Special attention will be given to the last problem. Some of the discussion is unique to research in managerial graphics, but much of it is generally applicable to experimental IS research.

Lack of underlying theory

As is true in other types of experimental IS research, computer graphics research lacks any adequately developed theoretical basis. The absence of theory contributes to the current state of inconclusive results in the literature because researchers lack a common ground for developing experimental hypotheses and interpreting results. The current trend in graphics research has been to perform one-shot, ad hoc studies without any significant effort to build upon the work of others and achieve a state of relatedness among studies. It is the authors' opinion that only through streams of directed research can investigators achieve an acceptable level of understanding of phenomena and ultimately formulate an underlying theory. The first step toward directed or "programmatic" research is the building of a framework that defines the boundary for research to be conducted. As Jenkins states, "...we need programs of research in MIS — programs of research based on a framework or frameworks, that provide a useful structure and focus on manageable subsets of the MIS field" [25, p. xi].
Jenkins further points out that research under directed programs should be (1) purposeful—not random, (2) cumulative and self-correcting—studies should build on the previous work, and (3) replicable—assumptions, tools, and procedures are clearly and explicitly communicated.

**Proliferation of measuring instruments**

Another factor that has contributed to weak and inconclusive results is the use of a great number of differing measuring instruments, many of which may have problems with reliability and validity. Research on presentation modes certainly suffers from this condition. For example, the research dealing with memory for information has used both recall (e.g., [31, 51]) and recognition measures (e.g., [20]). As a further illustration, "interpretation accuracy" has been measured by the accuracy with which data values displayed in different formats can be estimated [11, 12, 35], by the magnitude of errors found in interpretation tasks [5], by the percentage of correct responses to questions ranging from simple retrieval to complex decision making [48], by observing a trend in the presented data [42], and by comparing points or reading single points [43]. The use of different measures, even on the same construct variables, inevitably causes incomparable results and, therefore, leads to research labelled "conflicting."

Furthermore, the literature on presentation modes does not usually indicate whether instruments have been tested for adequate reliability and validity. In order to accumulate knowledge based on sound scientific inquiries, the relevance or validity of any instrument must be assured before relationships between measures of independent and dependent variables can be assessed. However, it must be acknowledged that the development of relevant and valid measurements is a very difficult process. To illustrate this point consider the measurement of a simple construct such as forecast accuracy in a situation in which subjects are asked to forecast three periods beyond a set of historical data. How is accuracy to be measured? By the average total forecast error? By the average of the absolute value of the forecast error? By the mean square error? Should any, a subset, or all of these measures be applied to each period forecast, all periods, or to only the first period (accuracy of forecasting the first period influences subsequent forecasts)? How does one compare a forecast based upon tabular data (which is precise) with graphical data (which, because of plotting inaccuracy, may be approximate)? Does one measure of forecast error do a better job in terms of "getting close" to the actual value? If so, what defines close? The problem is that this illustration is one of the simplest that we have confronted. Many of the measurement situations encountered can be significantly more complex and confusing.

Again, it is the authors' opinion that only through a program of research which is based upon standard tests can investigators hope for comparable studies. A set of measuring instruments, applicable and easily adaptable to a large number of experiments, must be developed. The construction of a set of good measuring instruments is a learning process which is lengthy, costly, and realistically feasible only when costs can be spread over several studies. Moreover, a research program provides a setting where the testing for reliability and validity of measuring instruments is almost a natural by-product of the studies themselves.

**Inappropriate research designs**

Research designs are considered inappropriate when (1) they do not address an important problem in the field, or (2) they are lacking in various forms of experimental control. In the area of graphics and information presentation some studies have restricted their measurements to constructs that have no direct link to decision maker productivity (e.g., user preference). Such inquiries do little to address the central issue of how to more effectively use computer graphics in IS decision making. Beyond the problem of irrelevant dependent variables, many experiments have been highly simplistic and included only one kind of independent variable. Studies which have examined two or more variables and their interactions are almost nonexistent. For example, at least 10 studies have compared bar charts with circle diagrams, and another 10 have
compared monochrome with color visuals. No study, however, has examined both graph type and color within the same experiment.

The second problem with research designs has been the lack of control. In particular, the control of variables which affect decision making, other than the presentation format, has been rudimentary in many experimental studies. Such studies raise questions about whether findings are less a function of presentation format than of other factors such as the situation, task content, personal background, etc. The condition of the task overwhelming the manipulation of presentation variables will be aptly demonstrated in the discussion of our series of experiments.

Two issues related to the lack of experimental control are: (1) non-equivalency of stimulus materials, and (2) subjects' greater familiarity with one presentation medium over another. Quality as well as content differences in the tabular and graphical presentation material probably have a significant impact on performance. The greater familiarity of subjects with tabular reports as contrasted to graphical reports may give the former a natural advantage which, in turn, may be moderated by the task [19].

To alleviate the research design problems addressed above, we suggest the following:

— Researchers should move away from univariate designs toward multivariate designs.
— Decision maker productivity, as opposed to viewer preference, needs further study as a dependent variable.
— To avoid confounding results, researchers should try to measure or control factors that are known from previous research to influence decision performance.
— Researchers should take proper actions to verify that the quality of their graphs are close to the quality of their tables or, as Ives [24] suggests, to at least make sure that both represent the highest quality possible.
— Researchers should clearly document the criteria used in generating the stimulus material, as well as any tradeoffs they have made that may have caused substantial content differences between presentation media.
— Training and learning effects of different media should be either manipulated or controlled so as to avoid unwanted familiarity biases with tabular versus graphical presentation.

Diversity of experimental tasks

The experimental task refers to the activity in which a subject is asked to participate in the course of an experiment. Task pertains not only to what the subject actually does, but also to the context, or surrounding environment in which the activity occurs. Graphics researchers have performed their studies in a multitude of task environments ranging from the employment of graphics in tracking military flight paths [21] to the use of graphics in aiding trust investment managers in day-to-day decisions [18]. The use of diverse and often unrelated and incomparable task situations makes the integration of findings across studies difficult because subjects' performance may be more a consequence of the task environment than of the use of graphics. In fact, there is a wide agreement that the characteristics of the task in which the subject is involved is a prime determinant of human decision making [16, 32, 41].

Several experiments have provided confirming evidence that the effectiveness of the display format is highly dependent on, or sensitive to the task at hand [1, 5, 7, 42, 43, 50, 54]. Thus, the type of presentation mode may have a relatively small effect on decision making performance compared to the task or task context. If this is true, then the results from many studies may be interpreted solely as a function of the task. Future research efforts will keep producing contradictory results unless researchers develop some type of taxonomy of tasks and start interpreting the results within the taxonomy. The development of a generally applicable taxonomy of tasks is a major research endeavor which can be achieved only through a long-term research program. A stream of studies is needed which consider characteristics of tasks such as complexity, task content, task difficulty, and task attributes (e.g., interpretation accuracy, trend spotting).
In our research we have confronted a related problem which is insufficient subject understanding of the experimental task. This problem leads to situations in which investigators are not able to determine what subjects are actually responding to in the experimental setting. Although this problem has been mentioned with regard to experimental IS studies using games, in which the decision environment can become highly complex (e.g., see [39]), it has not received the attention it deserves. It is our contention that much of the experimental IS research is plagued by this difficulty. What this means is that many results (or alternatively, non-results) are presented that are nothing more than the generation of random error, or "artifacts" of an experimental exercise. In these cases, the setting or task used for the research is not internally valid. This is a subject which deserves further elaboration.

**On the issue of internal validity**

Internal validity is in contrast only to external validity and not construct, content, or face validity. Cook and Campbell define internal validity as "the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause" [8, p. 37]. Thus, the lack of internal validity results in the inability to make any statements about cause and effect relationships and thereby invalidates the experiment.

In order to avoid internal validity problems, researchers should examine the cognitive processes used by some of the subjects in performing an experimental task. Task analysis involves separating task performance into its components so that subjects' psychological processes are revealed as they perform a task. Process tracing and debriefing of subjects are methods used to conduct task analysis. These procedures provide the investigator with vital information about the complexity and difficulty of a task and facilitate discovery of the methods and reasoning subjects use in a task. Unless a researcher is certain as to what subjects are responding to in the task the findings may be accidents, or may be due to factors other than the experimental manipulation, e.g., graphics.

**Summary**

We have argued for research on the managerial use of graphics that: (1) is programmatic and framework-based, (2) uses meaningful and tested measuring instruments, (3) uses appropriate research designs, (4) is based upon fully understood tasks, and (5) is internally valid. We will now set out to demonstrate how much effort is involved in attempting to achieve these objectives. We want to emphasize a disclaimer at this point — we have only begun to achieve these optimistic goals, and the work we will describe is only a step in what we believe is the right direction.

**The Experimental Program**

Recently a research group at the University of Minnesota began to study the managerial use of computer graphics. This group decided to follow a comprehensive programmatic approach in order to correct some of the problems from which other experimental IS research has suffered. For example, the Minnesota Experiments were framework-based but exhibited many of the other problems identified in the previous section. What follows is a description of a series of experiments which, over their duration, have illustrated how one ought to do this type of research rather than generated results showing how to effectively use managerial graphics. In particular, issues of internal validity and measurement have been stressed.

The research group, supported in part by a grant from the Society for Information Management (SIM), was interested in the relationships between graphical decision aids, task complexity, and decision making performance. Our first step was to choose an appropriate task environment and, within that environment, to define three levels of complexity.

**Task development and experimental design**

The issue of defining complexity turned out to be, in itself, a complex activity. Simon's conceptualization of complexity [44, 45] was adopted.
Simon identified two main contributors to complexity: (1) the number of elements in a system and (2) the degree and nature of the interactions among elements. For this experiment we selected only one of the factors (the number of elements in a system) to be manipulated to reflect levels of task complexity. We operationalized this construct in the task by manipulating the number of variables on which subjects received information from one experimental group to another.

The next step was to construct a case that would provide a task setting. First, an industry was selected based upon the criterion that products, operations, and markets of firms in that industry should be easy for subjects to understand. The business forms industry met this criterion. We focused on marketing operations since this area would be likely to benefit from the usage of graphics. Successful marketing management depends on identifying and examining trends and relationships of different marketing variables, and graphics are claimed to present trend and relationship information more effectively than other methods (e.g., [29, 30, 46]). To get a general setting for a case, the marketing operations of several business forms firms were studied. Other sources that provided vital information for the case were trade publications, annual and stock reports on business forms firms, and personal correspondence with a national trade association. The gathering of industry data aided in developing a realistic case. In addition, the case was written following the guidelines proposed for good case writing by Bennett and Chakravathy [3] and Reynolds [40].

The final case was three pages long and it described the industry, the company's current state, its products, and markets. The subject was to play the role of a consultant asked to help the CEO find the reason for falling profits at a time of increasing sales. The cause of the problem (unknown to the subjects) was the incorrect allocation of the salesforce among the firm's three markets. To be successful in the task, the subject was to determine that the salesforce was spending most of its time calling on customers in a market area with high sales revenues but severely declining profits; sales efforts should have been concentrated on a market that was smaller in terms of sales revenue but was highly profitable. We created several distractors that, we assumed, successful subjects would be able to reject as the problem after careful analysis. The distractors for each market area included such things as product pricing relative to the competition and advertising expenditures. Overall business conditions is an example of a distractor that applied to all three market areas. Within this task environment, three different situations were defined that varied by complexity level. The number of distractors determined the degree of task complexity (low = 3, medium = 6, high = 9).

The data for salesforce effort and distractors were presented in simple bar charts and grouped bar charts. The graphs were generated by a researcher who had no prior training in the techniques of presenting data in a graphical form. To help overcome this lack of experience, a "user-friendly" mainframe graphics software package was used for graphics generation. The graphs were carefully constructed to attempt to make the content and quality of different graph treatments as equivalent as possible. Nevertheless, problems occurred; for example, the grouped bar charts had somewhat "fuzzy" labels.

In addition to the graphs, questionnaires were constructed to gather information on (1) the backgrounds of subjects, (2) motivation of subjects, (3) subjects' satisfaction with the graphs, (4) the perceived complexity of the problem solving task, and (5) subjects' interpretation accuracy in reading the graphs.

The purpose of the first two measuring instruments, the background and post experiment questionnaires, was to control for the subject's marketing experience, managerial experience, educational level, previous familiarity with graphical tools, and motivation level in performing the experimental tasks. Lack of control of these variables has contributed to "no effect" results in several earlier studies reported in the literature (e.g., [26]).

Subjects' satisfaction with the graphical aids was appraised because earlier researchers had found the success of an MIS system to be correlated with user satisfaction (e.g., [37]). We tried to measure whether differences in satisfaction existed when different graph formats were
employed under varying task complexity levels. The "MIS Satisfaction Questionnaire" by Jenks [25] was modified for the experiment. The subjects were asked to rate their overall satisfaction with the graphs, as well as the readability and usefulness of the graphs, on seven-point scales. Jenkins' questionnaire was originally developed according to Nunnally's recommendations [33], and since the instrument had been validated before in an experimental setting, further validation was not considered necessary.

The fourth instrument was constructed to validate that the experimental tasks did in fact have different levels of complexity. Subjects were asked to rate the complexity of the problem finding task on a ten-point scale in comparison to a "base task," where the complexity of the base task was 5. The base task was a simple forecasting task that subjects undertook after the problem solving task.

The fifth instrument, an interpretation accuracy test, was developed to measure how well subjects could identify patterns, relationships, and exact data points from graphs. Its purpose was to appraise how accurately data was understood by subjects when displayed in a graphical form. Interpretation accuracy is a prerequisite to correct problem comprehension and improved decision quality. To assess the construct validity of this test, a sixth instrument, the Spatial Relations sections of the Differential Aptitude Tests (one of the most widely used and validated spatial relations tests) was given to each subject. We hypothesized that subjects whose performance was superior in the spatial aptitude test would also perform better than average on the interpretation accuracy test.

Preliminary testing

The experiment was pretested on two doctoral students, a faculty member, and an MIS practitioner. The subjects completed all the tasks in the experiment and also underwent a debriefing session in which they were asked to give a rationale for their answers in the problem finding task. The results were encouraging because three of the four subjects identified the correct rationale for finding the problem. The results of preliminary testing led us to believe that we had developed a good experimental task.

Experiment I

Design of Experiment I

The purpose of this experiment was to determine if the effectiveness in the use of graphical output depends upon the complexity of the task being performed by the decision maker. The study had two independent variables: (1) display format, and (2) task complexity. The display formats were simple bar charts and grouped bar charts. There were three levels of task complexity — low, medium, and high. There were four independent variables: (1) decision performance, (2) interpretation accuracy, (3) self-reports of satisfaction with the displays, and (4) decision confidence. The experimental hypotheses were as follows:

1. There will be differences in decision performance at varying task complexity levels.
2. There will be differences in decision performance when different graph formats are employed.
3. There will be differences in interpretation accuracy scores when different graph formats are employed.
4. There will be differences in satisfaction ratings when different graph formats are employed.
5. There will be differences in satisfaction ratings under varying task complexity levels.
6. There will be differences in decision confidence over varying levels of task complexity.
7. There will be differences in decision confidence when different graph formats are employed.
8. Subjects using grouped bar charts in the high complexity task will perform better in a problem solving task than subjects using simple bar charts.
9. Subjects using simple bar charts in the low complexity task will make better quality decisions than subjects using grouped bar charts.

The subjects were 63 graduate students at the University of Minnesota, 43 of whom were in the
MBA program with an MIS concentration. On the average, the students were 28 years old, had 41 months of full time, business-related work experience, and had successfully completed at least one marketing course.

The subjects participated on a voluntary basis and received course credit of 5% of the final grade in the course from which they were solicited for the study. All subjects were given an option of either completing a short class assignment or participating in the study. In addition to course credit, subjects received monetary prizes for good performance. The subject who performed best received $20, second best $10, and the three following performers, $5 each.

Procedure

During a two week period seven experimental sessions were conducted, with the number of subjects per session varying from 4 to 12. To discourage discussion about the experiment with subjects who were scheduled for later sessions, each participant was asked to sign a confidentiality agreement. Each subject completed a background questionnaire and then was given a maximum of 15 minutes for reading and answering questions related to the case description. The questions were asked to ensure that subjects had read and comprehended the case.

The problem finding task was given immediately after the case reading. Since time pressure in performing a task can affect the complexity of the task [52], it was very important to create a setting in which the high, medium, and low-complexity subject groups felt equally pressured with regard to time. A timing scheme for the problem-finding task was developed as a result of experience gained in pretesting. Subjects working on the low complexity task had 10 minutes; the medium complexity group had 15 minutes; and the high complexity group had 20 minutes. Within the allotted time, subjects studied the graphs and gave a written description of what they felt was the problem. They were then furnished with a list of potential problems. The list forced subjects to select a factor that they felt had contributed the most to the declining profitability of the business forms company. For the selected alternative, subjects also provided a subjective probability estimate on a scale from 0 to 1 of how confident they were about their choice.

The subjects next were asked to complete a 10-minute “base” forecasting task, followed by a questionnaire on the relative complexities of tasks, a satisfaction questionnaire, a 20-minute interpretation accuracy quiz, and a 25 minute spatial ability test. At the end of the session, subjects reported their motivation level in performing the tasks. Overall, the experiment took from one-and-a-half to two hours.

Results from experiment I

The analysis of experimental data did not reveal a consistent pattern of effects due to graphical and task complexity treatments. In fact, only the hypothesis on the differences in satisfaction ratings under varying task complexity levels was significant at a .05 level. A weak relationship was found between the performance of the high complexity task and the type of bar chart (see Table 1).

Problems resulting from experiment I

The results of the experiment raised concerns about the validity of the measures and tasks used. Beyond the nonsignificant experimental results, observations in the experimental sessions, subjects' comments, and the data collected concerning the validity of the measuring instruments and tasks all pointed to problems in the measuring instruments, research design, and the problem-finding task.

Our concerns about the measuring instruments focused on the validity of the interpretation accuracy quiz as well as the questionnaire on the complexity of the problem-finding task. The data collected for the validation of the interpretation accuracy test indicated a weak relationship with the spatial aptitude scores. This suggested the need for improving the test questions and possibly the quality of the graphs used as well. The experimental results did show that there were no significant differences between the low and high complexity groups in their perceptions of task complexity. This raised concerns as to whether the instrument designed to measure perceived complexity was invalid, or whether
Table 1. Major Hypothesis and Results for Experiment I

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. There will be differences in decision performance at varying task complexity levels.</td>
<td>no effect</td>
</tr>
<tr>
<td>2. There will be differences in decision performance when different graph formats are employed.</td>
<td>no effect</td>
</tr>
<tr>
<td>3. There will be differences in interpretation accuracy scores when different graph formats are employed.</td>
<td>no effect</td>
</tr>
<tr>
<td>4. There will be differences in satisfaction ratings when different graph formats are employed.</td>
<td>no effect</td>
</tr>
<tr>
<td>5. There will be differences in satisfaction ratings under varying task complexity levels.</td>
<td>*significant</td>
</tr>
<tr>
<td>6. There will be differences in decision confidence under varying levels of task complexity.</td>
<td>no effect</td>
</tr>
<tr>
<td>7. There will be differences in decision confidence when different graph formats are employed.</td>
<td>no effect</td>
</tr>
<tr>
<td>8. Subjects using grouped bar charts in the low complexity task will perform better in a problem solving task than subjects using simple bar charts.</td>
<td>**significant</td>
</tr>
<tr>
<td>9. Subjects using simple bar charts in the low complexity task will perform better in a problem solving task than subjects using grouped bar charts.</td>
<td>no effect</td>
</tr>
</tbody>
</table>

*Test for statistical significance < 0.05 level
**Test for statistical significance < 0.1 level

there had been ineffective operationalization of the levels of complexity in the experimental task.

Recall that the experiment involved the manipulation of three discrete levels of a complexity variable that, in theory, is continuous and can take on values ranging from low to high. The insignificant differences detected between the rankings of the high and low complexity groups (on perceived task complexity) led us to conclude that the task did not significantly vary with regard to task complexity. In other words, the levels of complexity in the three versions of the task were either all high or all low.

Other research design concerns were: (1) potential weak manipulation of stimulus complexity in the graphical formats, (2) inadequate monetary rewards, and (3) poor quality of stimulus materials. Since the display variable was a categorical variable no measure of how different the formats really were could be given. It was possible that the actual difference between simple bars and grouped bars is minor with regard to presenting data. Also the monetary rewards in the experiment may have been too low to motivate the subjects to perform well. Many of the subjects did not appear to be "trying hard" during the experimental sessions. Finally, the
quality of the graphs used as stimulus materials in the experiment may have been poor, thus contributing to insignificant findings. Although the researcher who prepared the graphs was knowledgeable about computer programming and took a course on how to use the graphics package, she had no formal training in graphic techniques. There were small layout problems and the labels in the grouped bar charts were fuzzy because of the low-resolution plotter used to generate hardcopy graphs.

In addition, the researchers were uncertain about what subjects were actually responding to in the problem-finding task. When the performance data was analyzed no consistent patterns in subjects' performance could be identified. In fact, subjects seemed to have guessed what the problem was. Perhaps the task was too difficult for subjects to handle or contained confusing or misleading information.

In sum, the researchers seriously questioned the validity of the task and the measures used in the experiment. Therefore, we decided to conduct another experiment which would focus on identifying remedies for our methodological problems, rather than on generating results on using graphics.

Experiment II

This study tested whether the results of Experiment I were caused by poor graphs or misleading or confusing information in the task. The research design was simplified by reducing the number of levels in the independent variables. Task complexity now had only two levels, and display format had one level. To help isolate the problems of poor quality and inappropriate layouts, 20 experimental subjects (all graduate students) received information in a tabular format. To examine whether the task contained misleading or missing information, subjects were asked to document what pieces of information they used from the tables and how they used the information in the task.

No significant performance difference was found between the high and low complexity tasks. The analysis of experimental data further indicated that the graphs had not been the problem in the first experiment. Subjects in the second experiment using tables performed as poorly as the participants in the first experiment using graphs. These results convinced us that the main problem was the fact that subjects just were not able to find the problem. However, what we did not know was whether the poor performance resulted from poor problem solving skills on the part of the subjects, or whether the task itself was misleading. Poor problem solving skills appeared as a likely reason because after the first two experiments we had realized that the actual task demands were much greater than originally intended. We suspected that even the low complexity treatment was a highly complex task because subjects had to relate variables to each other in order to be successful in the task.

Experiment III

The purpose of the third experiment was to determine whether there was misleading or confusing information in the task or whether the problem was poor problem solving skills on the part of the subjects. To accomplish this, 17 managers with an average of 10 years of managerial experience were used as subjects. It was assumed that if they could manage the task we could conclude that the task was valid, but too difficult for graduate students. To verify whether the managers successfully coped with the task, each subject was debriefed for 10 minutes at the end of the experimental session.

The results of the third experiment confirmed serious problems with the task. In particular, we found that managers identified the right problem, but often for the wrong reasons; and moreover, some subjects identified the wrong problem, although they correctly interpreted information portrayed through high resolution graphics. Obviously, the task was not providing the basis for answering our research question on the relationships among task, presentation format, and decision performance. As a result, a major revision of the task was undertaken.

Revision of the task

To revise the task we solicited the help of a marketing professor at the University of Min-
Methodological Issues in IS Research

Following his recommendations the case was revised to include more precise statements of the company's operations, and several marketing concepts were clarified that may have been confusing for subjects. In addition, the data was completely redeveloped resulting in a less complicated task in which the data patterns and relationships were easily detectable. In contrast to the previous task, the problem for declining profitability of the fictitious business forms company now clearly stands out (we think). Also, a graphic artist's recommendations were solicited on the graphics generated from the revised data, and small changes were made in the graph layouts.

The revised material is currently undergoing pretesting. Twenty-one graduate students have completed the revised task at this time, and 15 of them have detected the right problem in the problem-finding task.

Summary

We have gone through several experiments in searching and testing for a valid task and measurements. During this endeavor we have discovered that developing an effective task and accurate measurements is a lengthy and costly process that can best be described as an iteration of design and testing stages. After the initial design is completed, a series of pilot studies and refinements must follow. Through this multiple revision of the design, the internal validity of the study should increase, and moreover, the researchers should be more capable of estimating how valid the study is.

In the course of improving our experimental task environment we identified several problems. In Table 2, we have summarized these and other problems which we feel must be addressed when attempting to design valid studies. The problems are organized into 4 categories: (1) research strategy, (2) measuring instruments, (3) research designs, and (4) experimental tasks. In each case, suggestions are offered on how to alleviate or even avoid the problem. Many of the suggestions, however, require considerable "front-end" preparation (e.g., the search for a task should be a basis for a series of studies). Hence, it is important to incorporate suggestions early in the experimental design stage.

Another issue that has not received sufficient attention, but is an essential part of the experimental process, is the construction and conduct of pilot studies. Pilot experiments should not be quick-and-dirty preruns, performed a few days before the actual data collection. Rather, pilot studies should be carefully planned to ensure that a sufficient number of subjects has been recruited, that subjects are representative of the subject population to be used in the study, and that there is adequate time to revise the experiment and even run additional pilots prior to data collection. In particular, the selection of a convenient rather than a representative sample of pilot subjects is an easy trap in which to fall. This happened in our preliminary testing. We selected subjects who had shown interest in the study, and as a result we ended up with a group of high achievers that were likely to perform better than the average graduate student who participated in the experiment. This resulted in encouraging but faulty data on the validity of the experimental task. Also, extensive debriefings should be conducted, or protocols collected from the pilot subjects, to find out what subjects are doing when performing the tasks. This will help point out erratic components in decision making that may disguise the effect of treatments.

Even if all possible precautions are taken, a researcher may never design a perfect study because validity is a relative measure and therefore can only be estimated. In contrast to reliability, for which there are several statistical measures to help in the estimation process, there are no measures that exist for internal validity. Hence, we have to live with "rules of thumb" when trying to assess validity. Our goal is to adjust the task and the data until at least 50% of the experimental subjects solve the problem properly. Only then will we be able to investigate the impact of graphical aids on the decision process. The development of these rules of thumb is crucial in addressing a problem that we believe has been especially severe in experimental IS research using decision-making tasks. Take notice of how it became evident only after considerable checking, that the subjects were unable to perform the experimental task. Researchers must do considerable front
<table>
<thead>
<tr>
<th>Problems</th>
<th>Remedies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Strategy</td>
<td>Research Strategy</td>
</tr>
<tr>
<td>1. Piecemeal, unrelated research studies</td>
<td>1. Perform studies under a program of research that is at least framework driven, if not theory based.</td>
</tr>
<tr>
<td>Measuring Instruments</td>
<td>Measuring Instruments</td>
</tr>
<tr>
<td>1. Proliferation of incompatible measuring instruments</td>
<td>1. In the short run, modify and use, as much as possible, previously used and validated measuring instruments; develop your own only if absolutely necessary.</td>
</tr>
<tr>
<td>2. Proliferation of measuring instruments lacking in reliability and validity</td>
<td>2. For each modified or developed measuring instrument, collect data on (1) test-retest reliability across the samples of subjects that you further plan to use, (2) construct validity against other objective measures such as other validated instruments, (3) if the instrument is a modified version of the original, test the validity of the modified against the original.</td>
</tr>
<tr>
<td>Research Designs</td>
<td>Research Designs</td>
</tr>
<tr>
<td>1. Trivial and unimportant research designs</td>
<td>1. Avoid simplistic designs that focus on simple changes in graphic formats; investigate interaction effects of variables that affect decision performance or productivity.</td>
</tr>
<tr>
<td></td>
<td>Concentrate on the real issues and problems; use practitioners as a source for identifying real problems.</td>
</tr>
<tr>
<td></td>
<td>Question continuously whether a study is crucial and pragmatic.</td>
</tr>
<tr>
<td>2. Lack of control of extraneous variables that have resulted in confounding results</td>
<td>2. Control or measure and include into analysis: decision situation, subject experience, preference, task structure, task content, subject decision style, personality or any other variable that seems important to the research question under study.</td>
</tr>
<tr>
<td>3. Nonequivalent quality of tabular versus graphics presentation</td>
<td>3. Strive for equal resolution and layout; if not possible, use the highest quality available for each presentation.</td>
</tr>
<tr>
<td></td>
<td>Verify the quality with a graphics artist or equivalent expert.</td>
</tr>
<tr>
<td>4. Nonequivalent content of presentation stimuli</td>
<td>4. If equivalency cannot be achieved, then use following guidelines: include the same data with same precision into each presentation, clearly document the rationale for particular layouts, strive for best layout in each presentation media given a purpose of presentation, and get a graphic artist's opinion on the most effective layouts.</td>
</tr>
<tr>
<td>5. Biasness of results due to training and learning effects</td>
<td>5. Control or measure the familiarity of the subjects with the particular presentation media and layout format.</td>
</tr>
</tbody>
</table>
**Methodological Issues in IS Research**

For each individual task, assess the task complexity, task content, task difficulty and make comparisons between tasks only if known how the tasks differ on these dimensions.

**Conclusion**

The primary difficulty of experimental IS research is usually considered to be external validity [9]. Too often it is assumed that laboratory experiments, by their nature, are internally valid. This article challenges that assumption and attempts to convey the message that assuring internal validity in experimental research requires time and great care. To ensure a valid study, several pilot studies must be carefully conducted, and the issues of task and measurement validity must be addressed. We surmise that it may be a two or three year process to get the task and instrumentation correct, with a number of protocols and trial experiments conducted during this period. Outstanding discussions on internal validity by Campbell and Stanley [4], Cook and Campbell [8], and others [10, 23] are recommended for assisting researchers in avoiding internal validity and other design-related problems.

**References**


Methodological Issues in IS Research

25, Number 3, August 1979, pp. 787-798.


MIS Quarterly/June 1985 155


About the Authors

Sirkka Jarvenpaa is a Ph.D. candidate in MIS at the University of Minnesota. For the past two years she has been involved in research concerning the use of graphics in business decision making. Her dissertation examines the effects of work tasks and graphical information display on decision-making strategies and performance. The research is funded by the Society for Information Management.

Gary Dickson, Professor of MIS, is one of the co-founders of the MIS program and the MIS Research Center at the University of Minnesota. The author of over 50 articles and two books related to MIS, he is the founder and first Senior Editor of the MIS Quarterly.

Gerardine DeSanctis is an Assistant Professor of MIS at the University of Minnesota. Her current research interests are in graphical decision aids, group decision support, and implementation of information systems.