DAMAGED MERCHANDISE? A REVIEW OF EXPERIMENTS THAT COMPARE USABILITY EVALUATION METHODS

Wayne D. Gray & Marilyn C. Salzman

Human Factors & Applied Cognition

George Mason University

Fairfax, VA 22030

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*(3), 203-261.

HCI biography:

Wayne D. Gray is a cognitive scientist with an interest in how artifact design affects the cognition required to perform tasks. He has worked in government and industry; he currently heads the Human Factors and Applied Cognitive Program at George Mason University. Marilyn C. Salzman has worked as a usability engineer for industry; currently, she is a doctoral student in the Human Factors and Applied Cognitive Program at George Mason University. Her interests include human-computer interaction design and evaluation.

Send correspondence to:
Wayne D. Gray
George Mason University
m/s 3f5
Fairfax, VA 22030
(703) 993-1357
gray@gmu.edu

ABSTRACT

An interest in the design of interfaces has been a core topic for researchers and practitioners in the field of human-computer interaction (HCI); an interest in the design of experiments has not. To the extent that reliable and valid guidance for the former depends upon the results of the latter, it is necessary that researchers and practitioners understand how small features of an experimental design can cast large shadows over the results and conclusions that can be drawn. In this review we examine the design of five experiments that compared usability evaluation methods (UEMs). Each has had an important influence on HCI thought and practice. Unfortunately, our examination shows that small problems in the way these experiments were designed and conducted call into serious question what we thought we knew regarding the efficacy of various UEMs. If the influence of these experiments was trivial then such small problems could be safely ignored. Unfortunately, the outcomes of these experiments have been used to justify advice to practitioners regarding their choice of UEMs. Making such choices based upon misleading or erroneous claims can be detrimental--compromising the quality and integrity of the evaluation, incurring unnecessary costs, or undermining the practitioner's credibility within the design team. The experimental method is a potent vehicle that can help inform the choice of a UEM as well as help to address other HCI issues. However, to obtain the desired outcomes, close attention must be paid to experimental design.

TABLE OF CONTENTS

DAMAGED MERCHANDISE? A REVIEW OF EXPERIMENTS THAT COMPARE USABILITY EVALUATION METHODS

1. OVERVIEW

Usability is a core construct in human-computer interaction (HCI). Methods to evaluate the usability of various software packages have been of intense interest to HCI researchers and practitioners alike. Various usability evaluation methods (UEMs) have been created and promoted. The appeal of some UEMs rests upon common sense and/or the persuasiveness of proponents of that UEM. Others are based upon case studies or lessons learned and collected from various organizations. Finally, other UEMs have been promoted based upon the results of experimental studies designed to compare the effectiveness of two or more UEMs.

This review is limited to the latter sorts of arguments: experimental studies intended to yield objective and generalizable data regarding the utility of one or more UEM. As we will show below, the most influential of such experiments suffer from two basic problems. First, there are measurement issues. It is not clear that what is being compared across UEMs is their ability to assess usability. Although something is being measured, it is far from obvious that these measures really reflect sensitivity to usability. Second, there are design issues. The design of many of the experiments is such that neither the data they produce nor the conclusions drawn from the data are reliable or valid.

The implications we draw are of more than *academic* interest. They concern the entire HCI community. The parts of the review that examine the validity of assertions that one UEM is better than another may be primarily of interest to practitioners. The parts that discuss methodological and logical failings in experimental design may be of primary interest to researchers. However, the whole of the review is greater than the sum of its parts. Hence, this paper is not simply a review of experimental findings nor is it simply a methodology-oriented discussion of experimental failings.

Both authors are currently academics but formerly practitioners. We brought to our review the practitioner's disdain of hair-splitting academic arguments. However, as we warmed to our topic we found, time and again, numerous small problems in how the research was designed and conducted. If these small problems had only small effects on the interpretation of that research, we would have ignored them. Unfortunately, these small problems had large effects on what we could legitimately learn from the experiments. Cumulatively, they called into serious question what we thought we knew regarding the efficacy of various UEMs. We therefore ask our reader's indulgence in following us in the explanation of these small problems. This is not an exercise in hair-spitting. It is an attempt to convince you that much of what you thought you knew about UEMs is potentially misleading.

2. INTRODUCTION

UEMs are used to evaluate the interaction of the human with the computer for the purpose of identifying aspects of this interaction that can be improved to increase *usability*. They typically come into play sometime after needs assessment and before beta testing (see Olson & Moran, 1996, for a general discussion of where UEMs fit in the software development lifecycle). UEMs can be categorized as analytic or empirical. *Analytic-UEMs* include techniques such as Heuristic Evaluation (Nielsen & Molich, 1990), Cognitive Walkthrough (Lewis & Polson, 1992; Wharton, Rieman, Lewis & Polson, 1994), guidelines (e.g., Smith & Mosier, 1986), GOMS (Card, Moran & Newell, 1983; John & Kieras, 1996a; John & Kieras, 1996b), and others. *Empirical-UEMs* include a wide range of methods and procedures that are often referred to simply as user testing.

Our focus in this review is not on UEMs *per se* but on the studies that were intended as experimental manipulations to compare and contrast UEMs[1]. The purpose of these studies was to provide guidance to practitioners regarding the effectiveness of various UEMs. Our review tightly focuses on the design of the studies; what UEMs were used, how usability was measured, who the participants were, what the participants did, and so on. Our intent is threefold. First, we draw conclusions regarding whether the design of the study supports the claims that were made. Second, when the claims do not follow from the experimental manipulation, we identify the source of the problem. Third, over the body of studies reviewed, we uncover common problems and make suggestions for how experimental research in HCI can be improved.

In the cause of usability, doing something is almost always better than doing nothing. However, for HCI practitioners, making choices based upon misleading or erroneous claims can be detrimental -- compromising the quality and integrity of the evaluation, incurring unnecessary costs, or undermining the practitioner’s credibility within the design team. For UEMs to provide useful guidance to the design process, comparison studies must be conducted that delineate the tradeoffs -- the advantages and disadvantages -- of each method. Such experiments cannot be conducted quickly or easily. They require a substantial commitment of time and resources. Necessarily, all such experiments are limited in scope and these limits must be explicitly acknowledged. Such limits do not mean that guidance to practitioners cannot be forthcoming or that the experimental method does not have a role to play in formulating that guidance. If the power of the experimental approach is to be applied to illuminate the advantages and pitfalls of various UEMs, then broad brush experiments must be eschewed and a program of multiple, narrowly focused experiments must be embraced.

The plan for this paper is as follows. To delimit and focus the scope of our complaints, we begin by discussing the unique role of experiments among empirical methods. We then quickly review four well-known threats to validity in experimental studies (Cook & Campbell, 1979) and discuss a practice that seems widespread in the HCI literature: namely, *going beyond the data* to provide *advice* to practitioners. In the main section, we review five papers that have compared UEMs: Jeffries, Miller, Wharton, and Uyeda (1991); Karat, Campbell, and Fiegel (1992); Nielsen, (1992); Desurvire, Kondziela, and Atwood (1992); and Nielsen and Phillips (1993). All of these studies have been well cited; all were published in refereed conference proceedings (four in the prestigious SIGCHI proceedings and one in the well-regarded *People and Computers Conference* proceedings); and all have had enough time for fuller reports to appear in the literature (none have). Having provided a detailed introduction of threats to validity and having then used this framework to organize a detailed review of 5 studies, in the *Observations & Recommendations* section we suggest ways that future HCI experiments can be designed to avoid these threats.

3. THE UNIQUE ROLE AND BURDEN OF EXPERIMENTS IN EMPIRICAL STUDIES OF HUMAN-COMPUTER INTERACTION

Despite the plethora of alternatives, the traditional experimental approach continues to allure researchers. It is important to understand why. Simply put, “the unique purpose of experiments is to provide stronger tests of *causal* hypotheses than is permitted by other forms of research” (Cook & Campbell, 1979, p. 83).

A well-conducted, valid experiment permits us to make strong inferences regarding two important issues: (1) cause and effect and (2) generality. Experiments are conducted to determine the effect of some independent variable (e.g., type of UEM) on some dependent variable (e.g., the number of usability problems uncovered). Experiments permit us to go beyond noting correlation to inferring causality. For example, if using UEM-A, group-1 identifies more usability problems than group-2 (that used UEM-B), we can infer that it was the use of UEM-A, and not some other factor, that *caused* group-1 to find more problems (*effect*) than group-2.

Generality is as important to experimenters as causality. Is the effect found limited to the exact circumstances of the study or can it be generalized to other circumstances? If the study was conducted at NYNEX, can it be generalized to IBM? If the study used usability specialists with 3 years of experience, can it be generalized to specialists with 10 years of experience? Can the solutions to usability problems discovered by experimental methods be generalized to problems observed in usability testing?

The inference of cause and effect and the claim of generality is the essence of the experimental method and the reason for its continued attraction. However, few studies are generalizable across all times and places. Many have one or more limits in their claim to causality. When researchers become aware of problems and limits imposed by the conditions of their studies, it is their responsibility to call these limits to the attention of the reader and to explicitly circumscribe the claims made.

4. THREATS TO THE VALIDITY OF EXPERIMENTAL STUDIES.

What sorts of things should we look at to decide, first, whether a study was well done, and second, how far we can generalize the findings? These issues are often referred to as different types of validity. Although there are different breakdowns of validity, we consider the four discussed by Cook and Campbell (1979); statistical conclusion validity, internal validity, construct validity, and external validity. Additionally, we discuss the practice of using the discussion or conclusion section to go beyond what was investigated in the experiment to provide advice to practitioners. Each threat to validity is multifaceted and we focus below on those facets that are most relevant to HCI research. These facets will become major themes in our review of UEM studies.

4.1 Cause-effect Issues

Statistical conclusion validity and internal validity are concerned with drawing false positive or false negative conclusions. Succinctly put, statistical conclusion validity helps us establish whether the independent variable is related to the dependent variable. If there is a relationship, then the question for *internal validity* is whether we can conclude that the independent variable (the treatment) caused the observed change in the dependent variable (what we measured), or whether both variables are simply correlated and the observed changes were caused by a third, unnoted variable.

Statistical conclusion validity

Statistical conclusion validity is a realm well covered by statistics textbooks. The issues of particular concern for UEM studies include low statistical power, random heterogeneity of participants, and doing too many comparisons.

Low statistical power and random heterogeneity of participants might be regarded as two sides of the same coin. Low statistical power may cause true differences not to be noticed; random heterogeneity of participants may cause noticed differences not to be true. Potential solutions to these problems are to increase the number of participants per group or to consider group differences in the context of individual differences (variability).

Many UEM researchers use simple descriptive statistics (such as averages, percentages, and tallies) and tend to rely upon *eyeball tests*[2] to determine if apparent differences are real. Unfortunately, avoiding statistics does not avoid problems with statistical conclusion validity. When the power of the findings is too low to use a statistical test, the sample size may be too low to provide a stable estimate of an effect. In many cases of low sample size, effects due to the random heterogeneity of participants may be greater than the systematic effects of the treatment (e.g., type of UEM or type of expertise). Another way of phrasing this problem is to think of one or more of the participants as *Wildcards*; that is, people who are significantly better or worse than average and whose performance in the conditions of the study do not reflect the UEM but reflect their Wildcard status. The simplest solution to this *Wildcard effect*[3] is to randomly assign more usability interface specialists (UISs)[4] to each UEM treatment. The Wildcard effect is less likely to influence results when the Wildcard is one or two UISs in a group of 10 or 20 instead of one in a group of three. The checks on the Wildcard effect are statistics such as the t-test or ANOVA that compare the hypothesized systematic effect due to treatment (i.e., UEM) to the presumably random effect due to participant. If the treatment effect is enough bigger than the Wildcard effect then the difference between the two groups can be attributed to the UEMs and not to their Wildcards.

In addition to low power and random heterogeneity of variance (the Wildcard effect), the way in which comparisons are selected can pose a threat to validity. Experimental design handbooks such as Keppel and Saufley (1980), warn that "unplanned comparisons are considered 'opportunistic' in the sense that they can capitalize on chance factors" (pp. 140-141). They urge researchers to "take precautions against becoming overly 'zealous' in declaring that a particularly attractive difference" is real. The lure of interpreting these "particularly attractive differences" seems to be great in UEM experiments. In judging the potential problem posed by multiple comparisons, the important distinction is between comparisons that are chosen after examining the data versus those that follow from the logic of the experimental design.

Problems due to low power, random heterogeneity of participants, and unplanned comparisons can be controlled by the use of standard statistical techniques. Such techniques attempt to ensure that the random effect due to different participants is significantly less than the treatment effect due to the experimental manipulation. They also provide ways of doing multiple planned and unplanned comparisons while mitigating the capitalization on chance factors.

Experiments versus user testing. Statistical conclusion validity is more of an issue for those who would conduct experiments than for those involved in user testing. Since user testing assumes so many of the trappings of the experimental method it may seem reasonable to accept standards that are appropriate for user testing as appropriate for experimental research. However, this should not be done. A major distinction between the two is illustrated by the number of participants required to obtain meaningful data. *Experimental* results have shown that for *user testing*, only a few participants are needed to identify problems and even fewer participants are needed to identify severe problems (Virzi, 1992). (Though see Lewis, 1994, for a counter argument.) Similar studies were performed by Nielsen (Nielsen, 1992; Nielsen & Molich, 1990) and resulted in the recommendation to use 3-5 UISs for Heuristic Evaluation.

The distinction between standards appropriate for the usability lab versus those appropriate for research is well illustrated by the difference between the methods versus the recommendations of this series of studies. Support for the use of a handful of users or UISs in usability testing comes from experiments that used large numbers of participants (e.g., Virzi performed three studies with 12, 20, and 20 participants; Nielsen and Molich used 34 participants; and Nielsen used three groups of 31, 19, and 14 participants).

Internal validity

Statistical conclusion validity establishes that there are real differences between groups; internal validity concerns whether these differences are causal as opposed to correlational (there might be a third, unknown variable that is responsible for the changes). Unfortunately, there is no simple test for internal validity. There are, however, well-established issues researchers need to consider. Here we consider three: instrumentation, selection, and setting (see Cook & Campbell, 1979, for a fuller exposition).

Instrumentation. For UEM studies, instrumentation primarily concerns biases (covert or overt) in how human observers identify or rate the severity of usability problems. Comparing methods or groups is only valid if there is a way of rating the results that does not inappropriately favor one condition over the others.

In several of the UEM studies we review, evaluators were assigned to different UEMs and asked to identify usability problems using that UEM. This approach was threatened by instrumentation problems when either the evaluators or the experimenters were required to identify, classify, or rate the usability problems. For example, if, during the course of the study, the evaluators changed how they identified, classified, or rated problems, then an instrumentation problem existed. Factors that could have resulted in such changes are increased sensitivity to problems, increased experience with the system, etc. Another common flaw in the instrumentation of UEM studies is when problem categories defined by one UEM, for example, Heuristic Evaluation, were used by the experimenters to categorize problems found by another UEM, for example, user testing. In this case, the perspective of one UEM was likely to have biased the count or classification of problems found by the other UEM.

The issue is that *you find what you look for* and, conversely, *you seldom find what you are not looking for.* Although steps can be taken to prevent instrumentation problems, the scant discussion of this issue in the UEM literature raises rather than allays our concerns.

Selection. Selection is a threat “when an effect may be due to the difference between the kinds of people in one experimental group as opposed to another” (Cook & Campbell, 1979, p. 53). In our review we distinguish between general versus specific selection threats. A *general selection* threat refers to a characteristic of the participants that is not directly related to the manipulation of interest. A *specific selection* threat exists when the participants assigned to different groups are unequal in some characteristic (e.g., knowledge or experience) that is directly related to some aspect of the experimental procedures (and is not the intended manipulation).

Setting. The setting for an experiment may influence its outcomes. Indeed, an interesting research study might involve training consumers in, for example, Cognitive Walkthrough, and determining whether their evaluations of home software change as a function of applying the technique in a show room (before purchase), at home (after purchase), or in the usability lab. In this example, *setting* becomes the independent variable (i.e., something that is being manipulated to determine whether it exerts any significant effect on the outcome). In some of the studies we review, the setting covaried with UEM (treatment), type of participant (e.g., UIS vs. SWE), or both. In these cases, differences in setting is a threat to the study’s internal validity because it is impossible to determine whether the effect observed was obtained from the treatment, the setting, or the treatment-setting combination.

4.2 Generality Issues

If a study is internally and statistically valid, then we want to know whether the causal relationships found can be generalized to alternative measures of cause and effect as well as across different types of persons, settings, and times. Cook and Campbell (1979) refer to these issues as construct validity and external validity.

Construct validity

Construct validity divides neatly into two issues. Are the experimenters manipulating what they claim to be manipulating (the *causal construct*) and are they measuring what they claim to be measuring (the *effect construct*)? UEM studies have problems on both of these dimensions.

Causal construct validity

In the studies we reviewed, we found several types of threats to construct validity. The first threat is the most basic, as well as the most pervasive: operationalizing, or defining, the UEM. Did the way in which the experimenters conducted UEM-A correspond to the reader's understanding of UEM-A? Additional threats could arise from using only one way of applying a flexible UEM (mono-operation bias), applying the UEM to just one type of software (mono-method bias), or from the interaction of different treatments. Below, we discuss each of these issues in more detail.

Defining the UEM. The development and definition of UEMs has been a dynamic enterprise. In fact, all currently used analytic-UEMs have evolved rapidly over recent years. Therefore, it is understandable that the exact definition of any given UEM may have changed over time. Unfortunately, changing definitions while keeping the names the same makes it hard if not impossible to know what UEM is being manipulated and, hence, to compare outcomes from different UEM studies.

Analytic-UEMs such as walkthroughs, guidelines, and heuristic evaluations are the UEMs that suffer most from causal construct problems. To facilitate communication and comparison, we will use the terminology in Figure 1 which expands upon Olson and Moran's (1996) terminology.

INSERT FIGURE 1 ABOUT HERE

The two dimensions captured in Figure 1 are guidelines and scenarios. A scenario exists when evaluators are told to perform a given set of tasks or are asked to evaluate the steps of a task as they would be performed by the user (sometimes a flowchart is given, other times a listing is provided). Guidelines are broadly defined as any list of problems, or features, or attributes provided to evaluators for the purpose of determining whether any item from this list has been instantiated in the interface.

If experts are simply given an interface and told to evaluate it without being provided specific guidelines or specific scenarios, we call it an expert review (or simply a review if the evaluators are not experts). We reserve Heuristic Evaluation for Nielsen's "discount" technique (Nielsen, 1992; Nielsen, 1993; Nielsen, 1994a; Nielsen, 1994b; Nielsen & Molich, 1990) that among other things, provides evaluators with a short list of guidelines but no scenario. Having access to a long list of guidelines[5] but no scenario is referred to simply as guidelines.

When experts are simply given a scenario and told to use it in performing their evaluation, this is an expert walkthrough. A heuristic walkthrough provides evaluators with a short list of guidelines with which to identify problems found during the walkthrough. Similarly, a guidelines walkthrough provides evaluators with a long list of guidelines to use during the walkthrough. Finally, Cognitive Walkthrough is reserved for conducting a walkthrough using the techniques derived from CE+ theory (Lewis & Polson, 1992; Polson, Lewis, Rieman & Wharton, 1992; Wharton, Bradford, Jeffries & Franzke, 1992; Wharton et al., 1994).

From the perspective of causal construct validity, we cannot infer that UEM-A is better than UEM-B unless we are certain that the methods used by the experimenter were, in fact, representative of UEM-A and UEM-B. In the reviews that follow, when the terminology used by the researchers differs from the terminology we use in Figure 1, we will point out the difference but continue to use the terminology of the paper being reviewed.

Mono-operation and mono-method bias. Adopting Figure 1’s terminology does not ensure that different experiments have implemented the same UEM in the same way. Important variations may still exist. For example, in doing a Heuristic Evaluation, members of one group may conduct their evaluations independently and then combine results, whereas members of another may work together as a team. It is important to know (as researchers and practitioners) how changes in the way a UEM is carried out (i.e., *operations*) affect its ability to identify problems.

Mono-*method* bias is a complement to mono-operation bias. Just as there are many differences in how any one UEM can be used, there are many differences in the type of software that is to be evaluated. It is not obvious, for example, that a UEM that is good at finding problems with office automation software can be trusted to identify problems with real-time, safety-critical systems. Practitioners need to know what UEM works best for the type of software they are developing. Researchers need to understand whether and how usability problems vary with the type of software. Mono-method bias may lead us to draw false generalizations concerning both of these issues.

Interaction of different treatments: Confounding. In a few of the studies reviewed, two or more UEMs were used by the same set of participants. Such designs raise a threat to causal construct validity due to the possible *interaction of different treatments*. A threat exists because the experience gained by using UEM-A may affect the behavior (judgments or whatever) of participants while using UEM-B. For example, when using UEM-A participants are identifying problems as well as gaining familiarity with the software. Both these UEM-A outcomes (uncovering problems plus gaining familiarity) may be expected to feed into what the participants do and how they view the software later while using UEM-B. What is really being manipulated is not simply UEM-A versus UEM-B but UEM-A versus UEM-A/B; that is, the second treatment is not the pure form of UEM-B but is UEM-B *confounded* by UEM-A.

Effect construct validity: Intrinsic versus pay-off measures of usability

An important distinction for UEMs is the difference between *intrinsic* versus *pay-off* measures of usability (Scriven, 1977).

> If you want to evaluate a tool . . . say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you might just study the kind and speed of the cuts it makes in the hands of a good axeman. (p. 346)

Empirical-UEMs (i.e., most varieties of user testing) may measure performance directly. They are equivalent to Scriven's study of the "kind and speed of the cuts." Analytic-UEMs examine the interface or aspects of the interaction and infer usability problems. This is equivalent to Scriven's study of the bit, weight distribution, steel alloy, and grade of hickory. UEMs such as those of the GOMS family have demonstrated utility for relating intrinsic attributes of the interface to pay-offs (or costs) of using the interface (performance outcomes). For example, NGOMSL has been used to predict speed of learning (Bovair, Kieras & Polson, 1990; Kieras, 1997) and CPM-GOMS (Gray, John & Atwood, 1993, pp. 282-286) has pinpointed performance problems to specific features of the keyboard layout, screen layout, keying procedures, and system response times. For other UEMs, such as guidelines and Heuristic Evaluation, making this forward inference (from feature to pay-off) is not as tight or as obvious.

We find much confusion in the literature concerning the nature and role of the two types of UEMs. Analytic-UEMs examine intrinsic features and attempt to make predictions concerning pay-off performance. Empirical-UEMs typically attempt to measure pay-off performance directly (e.g., speed, number of errors, learning time, etc.). When an empirical-UEM is used to compare the usability of two different interfaces on some measure(s) of usability (e.g., time to complete a series of tasks) the results are clear and unambiguous: the faster system is the more usable (by that criterion for usability).

An opportunity for problems arises when the outcomes of empirical- and analytic-UEMs are viewed as equivalent; that is, when empirical-UEMs are used in an attempt to identify intrinsic features that caused the observed pay-off problem. Empirical-UEMs can identify problems, but care must be taken to isolate (e.g., Landauer, 1988) and identify the feature that caused the problem. None of the studies we reviewed report systematic ways of relating pay-off problems to intrinsic features, all apparently rely upon some form of expert judgment.

Problems of interpretation arise when the number of problems identified by one UEM is compared to those identified by another. When different techniques identify different problems, do the differences represent misses for one UEM or false alarms for the other? When comparing results of an analytic-UEM with an empirical-UEM, is a feature identified by intrinsic evaluation a usability problem even if it has no effect on performance?

We believe that *effect construct validity* is the single most important issue facing HCI researchers and practitioners. We will not dwell on this issue during our review of the 5 studies; however, we will return to this topic afterwards (see Section 6.1, p.51).

External validity

External validity concerns generalizing *to* particular target persons, settings, and times, and generalizing *across* types of persons, settings, and times (Cook & Campbell, 1979). The distinction is between generalizing to a population versus across subpopulations.

For example, Karat, et al. (1992) brought “together a group that had the composition of a development team (developers and architects, UI specialists, and appropriate support personnel) working with end users” (C-M Karat, personal communication, June 1, 1995). As intended by the experimenters, these results *generalize to* the heterogeneous population used by IBM development teams. However, readers who attempt to *generalize across* Karat et al.’s sample to groups composed of just UI specialists or just end users or just developers would be making an error of external validity. Again, the distinction is one of generalizing to a similarly heterogeneous population (okay) versus generalizing across the subpopulations (not okay).

Claims that exceed the scope of the *settings* and *persons* that the experiment can generalize *to* or *across* are said to lack external validity. The difference between the exact settings and persons used in the experiment and the wider range of settings and persons to which the experimenter seeks to generalize is a constant source of tension in reporting experimental results. When reporting results, researchers should attempt to balance grand claims against explicitly stated limitations. An example of such a qualification might be: “although we believe that UEM-A can be used by any UIS with no special training, we must note that the 12 UISs used in this study all had Ph.Ds. in Cognitive Psychology and had been involved, for the past five years, in a full-time effort to develop UEM-A.” Unfortunately, although broad claims are rampant, explicitly-stated limitations and caveats are rare.

4.3 Conclusion Validity

Are the claims made by the authors consistent with the results and/or do the claims follow from what was done? Our concern in this section is not with claims that are invalid due to one of the four Cook and Campbell validity problems. In the case of a problem with one of the first three types of validity (statistical conclusion validity, internal validity, or construct validity), presumably the research was designed to address that claim but ran into trouble for other reasons (that, unfortunately, were not noticed by the researchers). Likewise, for the fourth type of validity, external validity, presumably the claim made is an overgeneralization of a particular finding.

Our concern is with claims made in the discussion or conclusion of the study that were either (1) not investigated in the study or (2) contradicted by the results of the study. The former is *beyond the scope* of the study; whereas the latter is a *contradicted conclusion.* Although we can see no justification for contradicted conclusions, there are some interesting issues involved in claims that go beyond a study's scope.

For example, a reviewer (anonymous, personal communication, December 8, 1996) of an earlier version of this paper, lambasted us for our

> attempt to interpret everything said in the Conclusions or Discussion section of these papers as conclusions that logically follow from the analyses in the paper. Often statements made in such sections are comments or advice to others who might want to use these results.

There is a tradition in the human factors literature of providing advice to practitioners on issues related to, but not investigated in, an experiment. This tradition includes the clear and explicit separation of *experiment*-based claims from *experience*-based *advice.* Our complaint is not against experimenters who attempt to offer *good advice.* Rather, we are concerned with advice that is offered without the appropriate qualifications. Experience-based advice needs to be clearly and explicitly distinguished from experiment-based inference. Unless such care is taken, the advice may be understood as *research findings* rather than as the *researcher's opinion.*

4.4 Summary: Threats to Validity of Experimental Studies

Cook and Campbell (1979) discuss four threats to the validity of experimental studies. We have attempted to define these threats within the context of HCI research. In the next section, we examine the validity of five major studies of UEMs and discuss how validity problems threaten the conclusions one can draw based on these studies. We also note occasional contradicted conclusions, as well as experience-based advice that is not explicitly distinguished from experiment-based inference.

5. THREATS TO THE VALIDITY OF FIVE INFLUENTIAL UEM STUDIES

The body of this review examines five influential UEM studies. In order of publication, these are: Jeffries et al. (1991), Karat, et al. (1992), Nielsen (1992), Desurvire, et al. (1992), and Nielsen and Phillips (1993). Although it is arguable whether these are the *most* influential UEM studies, it seems beyond dispute that they have been *very* influential. Our arguments for influence are based upon informal surveys of the citation of these works in the proceedings of recent HCI conferences, journals, textbooks, and books. We recognize that many other noteworthy UEM studies have been reported in the literature. In fact, in an earlier version of this paper we included eleven extended reviews. However, for brevity, we have limited our discussion to a core set of reviews.[7]

Each review has three parts:

Overview: a description of goals of the experiment and a brief summary of the study's design and methodology.

Validity Issues: an examination of the validity problems that limit our acceptance of the claims made by the researchers.

Summary: our summary of the major problems with validity as well as what we believe can be safely concluded from *the study as conducted*.

INSERT FIGURE 2 ABOUT HERE

As a type of advanced organizer, we refer the reader to Figure 2 and Figure 3. For each study, Figure 2 indicates potential problems with how comparisons were made, the number of participants per group, and the type of statistics used (or the lack thereof). (Note that the individual entries in Figure 2 will be explained in detail as the study is discussed.) Figure 3 summarizes our judgment regarding each study’s most severe validity problems. In addition, we refer you to Appendices 1 through 5 for detailed lists of the claims made by the researchers for each study as well as the problems that threaten the validity of those claims.

INSERT FIGURE 3 ABOUT HERE

5.1 Jeffries, Miller, Wharton, and Uyeda (1991)

Overview

Jeffries et al. (1991) compared four UEMs that they called Heuristic Evaluation, Cognitive Walkthrough, guidelines, and user testing. By our classification, the first is an expert review rather than a Heuristic Evaluation; however, the others appear to be good exemplars of their categories (see Figure 1). All groups assessed the usability of HP-VUE™ "a visual interface to the UNIX operating system." There were four UISs in the Heuristic Evaluation group; one team of three SWEs in the guidelines group; one team of three SWEs in the Cognitive Walkthrough group; and six "regular PC users" who were not familiar with UNIX in the user testing group. The user testing was conducted by a human factors expert (UIS) practiced in user testing.

Jeffries et al.'s research goals were to determine (1) the interface problems the UEMs best detected, (2) the relative costs and benefits of each UEM, and (3) who (developers or UI specialists) could use the UEMs more effectively, all in a real-world setting. Based on the outcomes of the study, they drew inferences about each of these issues. However, a number of threats to validity limit the strength of their claims.

Statistical conclusion validity

Sample size. With 3 to 6 participants per group, this study suffered from low statistical power and its concomitant problem of random heterogeneity of participants; that is, just the sort of situation in which the Wildcard effect is likely to have occurred. With so few people per group, small variations in individual performance could have had a large influence on the stability of measures of group differences, making us unable to draw valid inferences about one UEM versus another.

There is a second and more subtle problem here as well. Participants in some groups worked as teams, making teams the unit rather than individual participants. Hence, Figure 2, which shows 4-3-3-6(UT) participants per group, actually provides a lenient count for sample size. A more accurate count would reflect the number of teams in a group, 4-1-1-1. The Heuristic Evaluation group had four teams (with one participant each) since each participant conducted his or her Heuristic Evaluation independently. Participants in the guidelines and Cognitive Walkthrough groups worked as teams, providing a sample size of one for each those groups. It also appears that we should think of the user testing condition as composed of one team since a single UIS conducted the user tests.

Statistics used. No statistical analyses were performed and the variability of individual performance was not accounted for. For example, the claim that "the [H]euristic [E]valuation technique as applied here produced the best results. It found the most problems, including more of the most serious ones, than did any other technique, and at the lowest cost" (p. 123) was based on a comparison of the total number of problems found by each group. Like this claim, all of Jeffries et al.'s claims were based on informal comparisons of totals, percentages, and means for very small sample sizes. Additionally, the large number of comparisons (see Figure 2) seems likely to have capitalized on chance factors.

Internal validity

Selection. Groups differed in the type of participant assigned to them, creating internal validity problems with selection. A *general selection* problem (see Internal validity p. 14) existed since UISs comprised the Heuristic Evaluation group, whereas the Cognitive Walkthrough and guidelines groups consisted of SWEs. Additionally, a *specific selection* problem existed in that two of the three SWEs in the guidelines group had extensive experience (40 and 20 hours) with the application (HP-VUE™) being evaluated but participants in the other groups did not. We have no way of knowing whether the finding that “the guidelines evaluation was the best of the four techniques at finding recurring and general problems” (p. 123) was due to the guidelines UEM or due to the experience with HP-VUE that members of the guidelines group had that was, apparently, not possessed by any member of any other group. These general and specific selection threats make it impossible to separate the influence of participant background from the utility of the technique.

Setting. Random irrelevancies in experimental setting presented an additional threat to internal validity. The Heuristic Evaluation group assessed HP-VUE at their own pace over a two week period and, presumably, at their own machines. The user testing group was given three hours of HP-VUE training followed by two hours of testing. Although the text is not clear, it appears that the guideline and Cognitive Walkthrough groups also completed evaluations in one sitting. Such extreme variations in setting might have affected group performance.

Instrumentation. To rate problem severity, Jeffries et al. (1991) provided their raters with the description of usability problems exactly as they had been written by the participants. No attempt was made to disguise the UEM group from which the problem statements came. In fact, the authors told us that problems identified by user testing were rated as more severe than other problems and that these ratings “may reflect a bias on the part of the raters” because “it was easy to tell which problems came from the usability test” (p. 122). If severity had to be judged by ratings (as opposed to a more objective method), then the raters should not have known (i.e., been blind to) which UEM group found what problem.

Construct validity

Causal. Because the meaning of the term *heuristic evaluation* has changed substantially since Jeffries et al. (1991) wrote their report, researchers have misinterpreted what this study has to say about UEMs. Heuristic evaluation has gone from being a primarily descriptive term to referring to a well-defined technique for evaluating usability. This problem was noted by Virzi et al. (1993) and was confirmed by Jeffries in an email exchange (R. Jeffries, personal communication, May 18, 1995). Jeffries et al.’s (1991) use of this term is faultless; indeed, it is exemplary as their description of their methods was such that a careful reader, such as Virzi, could map Jeffries et al.’s methods onto the changing UEM labels. However, as a reader, Virzi’s diligence is the exception, not the rule. For example, a clear misreading is Nielsen’s (1994b) citation of this study as showing that “Independent research has indeed confirmed that heuristic evaluation is a very efficient usability engineering method” p. 32. Such misreadings have lead to miscommunications and false conclusions by researchers and practitioners alike.

Effect: How problems were counted. Usability problems were counted for each UEM as they were found by participants in that UEM condition. There was no attempt to classify similar problems across UEMs[8]. Therefore, we do not know how much, if any, overlap there was in the problems or types of problems found by the different UEMs. Thus, if we do not know whether the problems found using Cognitive Walkthrough and guidelines were actually the same kinds of problems, the conclusion that “the [C]ognitive [W]alkthrough technique was roughly comparable to guidelines” (p. 123-124) may be misleading. Likewise, if we do not know the extent to which problems found using the different UEMs were unique, the conclusion that “the [H]euristic [E]valuation technique as applied here produced the best results...” (p. 123) is not as informative as it might otherwise be.

External validity

A very particular combination of settings, evaluators, and UEMs was used in this study. This combination would make it impossible to generalize the results to other persons, places, or variations on UEM operations.

Conclusion validity

As discussed in our overview of Jeffries, et al., these authors cite as one of their goals, the determination of how expertise affects performance. They follow up on this in their conclusions, stating that Heuristic Evaluation is dependent upon “having access to several people with the knowledge and experience necessary to apply the technique” (p. 123). However, the study was not designed to examine either the number of people necessary for an Heuristic Evaluation nor the knowledge and experience required for conducting the evaluation. Thus, this conclusion goes beyond the data.

Summary of the review of Jeffries, et al. (1991)

If Jeffries, et al. (1991) had been cast as a case study (and appropriate changes made throughout), the paper would have provided a snapshot of the trade-offs facing Hewlett-Packard in deciding how to do usability analyses in the late 1980s. Unfortunately, the work was presented as an experimental comparison of four UEMs and several misleading conclusions were drawn. Those regarding one UEM versus another were weak because of low power and lack of statistics, as well as uncontrolled differences (i.e., setting and selection) among groups. Claims made about the types of problems found by each UEM are problematic for the same reasons. In addition, there are construct validity issues with how problems were counted. Finally, conclusions regarding the strengths of different types of evaluators went beyond the scope of the study. Overall, the design and scope of the study did not support the inferences made regarding cause and effect or generality of the results. (See Figure A1 for a detailed list of claims and the validity problems that weaken them.)

5.2 Karat, Campbell, and Fiegel (1992)

Overview

Karat, et al. (1992) compared user testing with a walkthrough technique that combined scenarios with guidelines (heuristic walkthrough by our definition). Forty-eight participants were drawn from a participant pool consisting of users, developers, and UISs. They were assigned randomly to three conditions: user testing (two groups of six individuals), individual walkthrough (two groups of six individuals), and team walkthrough (two groups with six teams of two individuals per team). One group in each condition evaluated one integrated office system (text, spreadsheet, and graphics), whereas the second group evaluated a second integrated office system. (See Figure 4.) During a three-hour session, participants used the technique to which they had been assigned to learn about the system, freely explore the system, work through nine scenarios, and complete a questionnaire.

INSERT FIGURE 4 ABOUT HERE

Karat et al. used the outcomes of this study (1) to compare the number of problems found, problem severity, and resources required by each UEM, (2) to determine if differences between UEMs generalized across systems, and (3) to examine how the characteristics of walkthroughs influenced effectiveness (e.g., the effectiveness of individual vs. team evaluations, the effect of evaluator expertise, the value of scenarios). Although the study's design was fairly strong, a few validity problems undermined the soundness of some of the authors' conclusions.

Statistical conclusion validity

Sample size. Recognizing that, in their team walkthrough condition, the team was the unit of analysis and not the number of individuals per team, Karat et al. went to some lengths to ensure that each group had six participants (6 teams or 6 individuals, see Figure 4). Although six participants per condition was not a large number, 12 participants for each UEM condition and 18 per system (see Figure 4) may have provided enough power to guard against the Wildcard effect. However, as we discuss in the next section, the authors' analyses failed to take advantage of this strength in the study's design.

Data analysis. The majority of the comparisons reported followed from the design of the study and clearly were not opportunistic efforts to capitalize on chance factors (see Section 0). However, the comparisons made and conclusions drawn rested upon simple descriptive statistics (e.g., averages and percentages) or χ^2 analyses that did not take into account the variability of participants. With 18 participants per system, (see Figure 4) tests that considered participant variability should have been performed. Without the support of statistics to compare group differences against participant (subject) variability we should not infer that differences in group performance were due to treatment rather than participants; that is, the Wildcard effect has not been ruled out.

Internal validity

A possible *general selection* problem exists. Two UISs administered user tests; in contrast, a combination of users, UISs, and SWEs completed the walkthroughs. This difference may have contributed to the finding that user testing was better than walkthroughs.

Construct validity

Dealing with mono-operation and mono-method bias. Karat et al. directly compared two ways (or sets of operations) of conducting walkthroughs: individual versus team. This comparison was a noteworthy attempt to move from simplistic statements (“UEM-A is good”) and offer more useful guidance (“if you have three UISs, have them use UEM-A as a team”). Also noteworthy was the use of two different business software packages in the same experiment. This practice probably reduced mono-method bias and facilitated the generalizability of their findings. However, their manipulation would have been more useful if the authors had been able to characterize the nature of the differences between the two systems without giving away proprietary information. As is, readers are provided with no guidance in mapping from the particular systems used by Karat et al. to characteristics of systems that they might wish to evaluate.

Confounding. Karat et al. claimed that “All walkthrough groups favored the use of scenarios over self-guided exploration in identifying usability problems. This evidence supports the use of a set of rich scenarios developed in consultation with end users” (p. 403). Without careful qualification, the design of the study does not permit us to conclude anything about the value of scenarios. First, all participants completed self-guided exploration first and scenarios second. Hence, experience gained during the self-guided phase was available to these groups during the scenario phase. This uncontrolled order effect of self-guided exploration on scenarios threatened the causal construct validity of their claim by confounding the treatments. Second, although participants liked the scenarios, no evidence was presented to suggest that the scenarios helped them find more problems. Thus, this claim goes beyond the scope of the study.

Effect construct validity. To support their conclusion that “Team walkthroughs achieved better results than individual walkthroughs in some areas”(p. 403) Karat et al. provided a χ^2 analysis of the number of *problem tokens* found. This analysis showed a statistically significant difference favoring team walkthrough over individual walkthrough. However, Karat et al. classified problems in four ways: tokens, types, problem areas, and *significant problem areas*; it appears that these other measures were not necessarily consistent with the authors’ interpretation. For example, for problem areas, the means for the team and individual walkthrough groups were identical. In terms of significant problem areas, teams found more problems than individuals for system 1 (mean of 3.83/team vs. 3.00/individual) but not for system 2 (mean of 2.33/team vs. 2.83/individual). Hence, the conclusion favoring team walkthroughs over individual walkthroughs depends upon the usability measure that was chosen; that is, had the analysis been based on the number of problem areas or significant problem areas, it would have told a different story.

External validity

The external validity of the findings must be tempered by two considerations. First, the development team members (SWEs) used here were not members of the team that developed the product. Thus the finding that “users and development team members can complete usability walkthroughs with relative success” may not generalize to SWEs that are asked to evaluate software that they have developed.

Second, the authors can generalize to mixed teams of GUI users, UISs, and SWEs; however, readers should be careful about generalizing these results *across* the population (see discussion of External Validity, p. 22) to teams composed of just GUI users, or just developers, or just UISs. For example, “relative success” may be due to the synergy of a mixed team, solely due to the SWEs, or solely due to the GUI users.

Conclusion validity

Unfortunately, many of authors’ conclusions are either beyond the scope of their study or are contradicted by their own data. For example, the authors claimed that studies by Jeffries, et al. (1991), Desurvire, et al. (1991), as well as their own study “provide strong support for the value of [user interface] expertise” (p. 403). This claim goes well beyond the data. User interface expertise was not varied in Jeffries, et al. (1991), nor was it varied here[9]. Since user interface expertise was not manipulated in this study, this is conclusion is beyond the scope of the study. In a discussion of the significant problem areas identified across methods, Karat et al. claim that “These methods [user testing and walkthroughs] are complementary and yield different results; they act as different types of sieves in identifying usability problems” (p. 403). This is a contradicted conclusion. Based on the data presented in Karat et al.’s Table 4, the UEMs did not yield different types of data. Across the two systems, the user testing group identified 21 problems (called significant problem areas by the authors) not found by the team or individual walkthrough groups. The two walkthrough conditions together found only 3 problems not found by the user testing group. Thus, walkthroughs did not catch many problems missed by user testing.

Summary of the review of Karat, et al., (1992).

This study handled most of the threats to internal validity well and thereby provides a model of how experimental research can be conducted within a corporate environment. The mixed nature of the groups limits the generalization (external validity) of their findings. Given the evaluation philosophy at IBM when this research was conducted, this limit is both reasonable and fair; however, the authors might have stressed this limit in their discussion and conclusion sections more than they did. There also are minor construct validity problems concerning several issues. These problems raise concerns with how the study was interpreted, but none of the problems can be considered a fatal flaw. The main failing of this study was with statistical conclusion validity. Few statistical tests were reported and those that were reported failed to control for the Wildcard effect. Hence, although the results regarding the superiority of user testing to walkthroughs may be interesting and suggestive, they may not be generalizable beyond this study’s testing conditions. Additionally, some of the claims made about the study have problems with conclusion validity. Thus, several of the findings should be considered with caution. (See Figure A2 for a detailed list of claims and validity problems.)

5.3 Nielsen (1992)

Overview

In this paper, Nielsen described (1) a study in which he examined whether the probability of finding usability problems increased with usability expertise as well as domain (voice response systems) expertise and (2) a study where he classified outcomes for six Heuristic Evaluations of different user interfaces along several dimensions. In the expertise study, Nielsen had three groups complete a Heuristic Evaluation of a printed dialogue (as opposed to a running system). Groups consisted of 31 computer science students (novices) who had completed their first programming course; 19 UISs (single experts) who had "graduate degrees and/or several years of job experience in the usability area" but with no special expertise in voice response systems; finally, 14 double experts who "had expertise in user interface issues as well as voice response systems." Nielsen interpreted the results of his study to provide advice regarding how expertise effects the types of problems found by Heuristic Evaluation.

In the classification study, Nielsen characterized problems found in six different Heuristic Evaluations along the following dimensions: severity (two levels), heuristic (nine heuristics), location (four locations), and type of system (two types). The results of this classification were presented in two forms: the proportion of problems falling into a category per evaluator (mean proportion) and the total proportion of problems aggregated across evaluators. Using these kinds of classifications, he attempted to describe Heuristic Evaluation's strengths and weaknesses in facilitating problem identification. However, problems with assumptions made about the effect construct as well as the approach used to analyze the data severely weakened the validity of these studies, as well as many of Neilsen's claims.

Statistical conclusion validity: Sample size and data analysis

In the expertise study, the large number of participants per group (31 vs. 19 vs. 14) should have mitigated the Wildcard effect and should have warranted statistical tests. Unfortunately, none were reported. Although statistical tests were not conducted, possible differences were noted and discussed as if they were real. For example, the conclusion that "usability specialists with expertise in the specific kind of interface being evaluated [double experts] did much better than regular usability specialists without such expertise [single experts], especially with regard to certain usability problems that were unique to that kind of interface" (p. 380) relies on a comparison of group means.

In his analysis of Heuristic Evaluations, threats to statistical conclusion validity are even more severe. Not only was the variability of the data unaccounted for when comparisons were made, but the comparisons were selected from a large body of potential comparisons, substantially increasing the probability that apparent differences were due to chance. To illustrate how this invalidated conclusions based on these data, consider the following claim: "Problems with the lack of clearly marked exits are harder to find than problems violating other heuristics" (p. 380). If we isolate the heuristic comparisons, 306 were possible.[10] Thus, this claim was based on 1 seeming difference among a possible 306 differences.

Overall, Nielsen’s Table 2 provided a possible 3546 comparisons by system and another 374 if the data were aggregated by type of prototype.[11] Picking a few particularly attractive differences out of this sea of potential comparisons, as well as failing to consider the stability of the data (their variability) upon which these differences were based, was likely to have capitalized on chance factors.

Internal validity

Unfortunately, the paper does not provide the details necessary to assess thoroughly either study’s internal validity (e.g., general or specific selection threats). For example, in the expertise study, we are not told how long each evaluator spent on the task or whether time on task was limited by the experimenter or up to each evaluator. We are also not told whether the evaluation was completed individually or in groups or anything else about the methodology or conditions of the study. Nevertheless, an instrumentation problem is apparent in both studies.

Instrumentation. In the expertise study, classification of usability problems into minor problems or major problems rests upon “a considered judgment” (p.379). Whereas such an informal basis may suffice for many situations, it does not support a claim as complex as "Major usability problems have a higher probability than minor problems of being found in a heuristic evaluation, but about twice as many minor problems are found in absolute numbers" (p. 380). A similar problem was inherent in the classification study, as classifications were made in much the same way. (Note that whereas *what* measures were made is an effect construct validity issue, *how* these measures were made is an internal validity problem with instrumentation.)

Construct validity

In the expertise study, if we ignore potential problems with statistical conclusion validity then we might conclude that novices named fewer potential problems than the single experts, who, in turn, named fewer potential problems than double experts. Note that the phrase we use is “named fewer potential problems” and not “found fewer actual problems.” The *construct validity of effect* was a major weakness in this study. Heuristic Evaluation was not compared to user testing or to any other dependent variable. Anything that an evaluator named was counted as an actual problem. There is no way to be sure that the named problems (intrinsic features) would have corresponded to real problems (pay-off measures of performance).

Conclusion validity

If potential problems with statistical conclusion validity are ignored then a modest claim such as *experts named more potential problems than non-experts* would have been supported by the data. However, the broader claim that "usability specialists were much better than those without usability expertise at finding usability problems by heuristic evaluation" (p. 380) was not. There are two issues here. First, no attempt was made in this study to determine if the *named problems* were *actual problems.* It seems likely that there was more than one false alarm (calling something a problem when it is not) or miss (not finding a problem when there is one). Second, whereas the effect of expertise on an evaluator’s ability to name problems was clear, the increment contributed by Heuristic Evaluation was not. How many of the named problems were found by expert judgment? How many more named problems did Heuristic Evaluation contribute? As these questions were not addressed in this study, conclusions about them cannot be drawn. A similar weakness existed in the classification study. It was intended to demonstrate the strengths and weaknesses of Heuristic Evaluation by categorizing problems found through that technique. However, it was never established that the problems being categorized were in fact real usability problems - ones that affected performance; nor was it established that Heuristic Evaluation was responsible for revealing those problems.

Summary of the review of Nielsen (1992)

The author attempted to tackle some critical issues and was creative in finding ways to address those issues based on limited data. However, some of his answers may be misleading. Assumptions made about the effect construct as well as the approach used to analyze the data severely weakened the validity of Nielsen’s conclusions. His conclusions assumed that named problems were actual problems and that differences found by opportunistic comparisons and tested only by eyeball statistics were real. Therefore, we need to be careful in how we interpret his advice on expertise and the strengths and weaknesses of Heuristic Evaluation. (See Figure A3 for a detailed list of claims and validity problems.)

5.4 Desurvire, Kondziela, and Atwood (1992)

Overview

Desurvire, et al. (1992) compared the effectiveness of three types of evaluators (UISs [experts], SWEs, and non-experts) on two analytic-UEMs: Heuristic Evaluation and Cognitive Walkthrough. Additionally, all UEM conditions were compared with user testing. The Heuristic Evaluation (heuristic walkthrough in our terminology) and Cognitive Walkthrough groups used “paper flow-charts organized by task” to complete six tasks. The user testing group used a prototype of the interface (H. Desurvire, personal communication, October 20, 1995).

The authors' research goals were to determine the value of expertise and to assess the relative strengths and weaknesses of the analytic UEMs compared to user testing. They drew a number of conclusions relating to each of these goals. Unfortunately, weaknesses in the study's design and the analyses used are cause for questioning the validity of all of these claims.

Statistical conclusion validity: Sample size

The user testing group had 18 participants; the six analytic-UEM groups had three (3) participants each. These participants were distributed as shown in Figure 5. The small number of participants and their distribution among groups raise concerns with statistical conclusion validity as well as internal validity.

INSERT FIGURE 5 ABOUT HERE

Desurvire et al. (1992) attempted to justify the use of three participants per group by citing Nielsen and Molich (1990) and Nielsen (1992), who recommended using three evaluators when conducting a Heuristic Evaluation. Unfortunately, as discussed in section 0 (*Statistical conclusion validity* p.12), this confusion of the standards appropriate for the usability laboratory with standards appropriate for experimentation threatened the statistical conclusion validity of the experiment. The use of too few participants in an experiment results in low power, unstable estimates of group performance, and a tendency for the Wildcard effect.

Additionally, the authors were overly zealous in interpreting almost every difference between two numbers as *real* (we count 57 such claims).

As an example, consider Desurvire, et al.'s (1992) claim that "[UISs] in the Heuristic Evaluation condition named almost twice as many problems that caused task failure or were of minor annoyance in the laboratory, than [UISs] in the cognitive condition". Because this claim was based on a comparison of *totals* for three participants per group, this statement is misleading. Errors were classified according to problem severity: "minor annoyance," "caused error," and "caused task failure" (see Desurvire, et al.'s Table 2). The Heuristic Evaluation group found a grand total of 4 "minor annoyance" problems whereas the Cognitive Walkthrough group found 2. Likewise, for "caused task failure" problems, the Heuristic Evaluation group found 5 problems and the Cognitive Walkthrough group found 3 ("almost twice as many"). *These numbers do not represent the average number of errors found per evaluator but the total number of errors found per group*. Although the arithmetic is indisputable - four problems is twice as many as two - the meaningfulness and reliability of these claims are questionable.

Internal validity: Selection

A specific selection threat exists in that the same three SWEs evaluated the same six tasks using both Heuristic Evaluation and Cognitive Walkthrough. Hence, in one of these two conditions (we are not told which condition the SWEs completed first) the SWE group had more experience with the task than the UIS group did, making any comparisons among groups difficult to interpret. The second UEM performed by the SWEs had to have been affected by their increased familiarity with the system, with the six tasks, as well as by their memory of the problems found by the first UEM. Thus, it is hard to assess the validity of the claim that “...there were no differences between methods for the [SWEs]”

Construct validity and external validity

The above problem with the internal validity of selection also results in a confounding of treatment conditions. This problem, combined with the previously noted problems with statistical conclusion validity, undermine the possibility of anything but a limited generalization of these findings.

Summary of the review of Desurvire, et al., (1992)

It is evident that the authors were eager to share their research and were careful to identify many important questions. Unfortunately, they failed to recognize the limitations of their study and based many strongly worded conclusions upon scant data. The prerequisites for an experimental study, statistical conclusion validity and internal validity, were severely lacking. Due to this lack of prerequisites, we do not offer a detailed evaluation of the study’s construct or external validity. We believe that there is nothing that can be safely concluded regarding UEMs or expertise based upon this study (see Figure 2 and Figure 3). (See Figure A4 for a detailed list of claims and validity problems.)

5.5 Nielsen and Phillips (1993)

Overview

Nielsen and Phillips (1993) compared (1) performance time estimates and (2) the costs of four analytic-UEMs and user testing. The four analytic-UEMs were: Cold, Warm, and Hot heuristic estimates[12]; and keystroke level modeling GOMS[13] (KLM). They asked participants in each of the four analytic-UEM conditions to use that UEM to estimate the time it would take users to perform tasks using two interfaces (Dialog Box vs. Pop-Up Menu) to the same database. Each participant in the Cold and KLM groups independently evaluated the systems based on written specifications only. The Warm group based estimates on a prototype of the Dialog Box interface and written specifications of the Pop-up Menu interface. The Hot group used running prototypes of both systems when producing their estimates. Time estimates obtained from the four analytic-UEM groups were compared to actual times for a user test group.

Based on the outcomes of this study, the authors drew conclusions about the relative effectiveness and costs of the UEMs, as well as their reliability. They also made claims about each UEM's ability to support evaluators in making absolute and relative estimates of system usability. However, weaknesses in their methodology and analyses threatened the validity of many of their claims.

Statistical conclusion validity

Sample size. Groups appear to have been sufficiently large to avoid the Wildcard effect. There were 12 Cold, 10 Warm, 15 Hot, 19 KLM evaluators; and 20 participants in the user test.

Appropriateness of statistics. Formal statistical tests were not used; that is, conclusions in this study were based on an inspection of the means, without considering variability. In fact, the data were highly variable and the authors' own estimates (see their Table 2) show that they would have needed many more participants to establish reliable measures of group performance. For example, the claim that "Heuristic estimates were better in the hot condition where estimators had access to running versions of the two interfaces, than in the cold condition based on specifications only" (p. 221) is based an inspection of means with very high variability.

Independence of measures. The authors assessed the cost of each UEM by comparing the average time to complete evaluations. Unfortunately, as discussed below, the Warm group had prior experience with the system. Recognizing that this prior experience would have distorted the estimates of how long it took the Warm group to perform the task, the authors used the average of the Cold and Hot group evaluation completion times to assign a completion time to the Warm group. This assignment of time to the Warm group violated the statistical conclusion validity assumption of independence of measures.

Internal validity

Selection. A *general selection* problem existed in that the expertise and background of the KLM and the heuristic estimation groups were not equivalent. "For all three heuristic estimation conditions, the evaluators were usability specialists with an average of nine years of usability experience" (Nielsen and Phillips, p. 217). For the KLM condition the evaluators were undergraduates doing their second KLM assignment at Lewis & Clark College. Far from experts, half were psychology majors and the rest were English, art, physics, and history majors (E. Nilsen, personal communication, April 18, 1995). (Note that some of these observations have been made by John, 1994.) Thus, any comparisons between the KLM and heuristic estimation groups (e.g. "GOMS and heuristic estimates were about equal..." p. 221) are questionable.

Additionally, a *specific selection* threat arose from differences between the heuristic estimation groups in prior experience with the system being evaluated.

Warm group UISs had completed a Heuristic Evaluation, rated severity problems, and heard a complete explanation of the full system prior to participating in this study. All in all, they spent about 2.5 hours on such activities. The other groups had no prior exposure to the system. The effect of this prior experience on heuristic estimation is unknown and makes any comparison involving the Warm group suspect.

Settings. Differences in setting were also likely threats to validity. Conducting evaluations in the Bellcore workplace versus in a Lewis & Clark College dorm room must have been very different experiences that may have affected the findings in indeterminable ways. Again, this would have affected the validity of comparisons involving the KLM group.

Instrumentation. The college students in the KLM group were not consistent in how they estimated the time it took them to do the evaluation. Some of the students included time spent to “write a memo to their fictional manager explaining the approach and their recommendations for the new application” and some did not (E. Nilsen, personal communication, April 18, 1995). This is an instrumentation problem that threatens internal validity.

Construct validity and external validity

The problems noted above with internal validity undermine the possibility of anything but a limited generalization of these findings.

Conclusion validity

Nielsen and Phillips claimed that, “Performance estimates from both heuristic estimation and GOMS analyses are highly variable” (p. 221). This claim does not appear to be supported by their data. Although outcomes for heuristic estimation do appear to be highly variable, outcomes for KLM analyses do not. If the authors had in mind an implicit comparison with user testing, then the variability of KLM estimates (with novice, KLM analysts) compared well with the variability of user testing. From their Table 1 we see that the "Standard Deviations as % of Means" for the cold, warm, and hot heuristic estimate conditions were: 108%, 75%, 52%. In contrast, for KLM (shown in the table as "GOMS") and User Testing the ratios were 19% and 17% (see John, 1994 for a replication of the KLM results). Although it is clear that the heuristic estimations of experts in all three conditions were highly variable, the estimates from KLM novices were not. Hence, the part of this claim that applies to the KLM estimates is contradicted by the data.

Summary of the review of Nielsen and Phillips (1993)

The authors used enough participants to estimate variability and provided estimates of this variability in their presentation of the data. In fact, these variability estimates strongly support a point we have been trying to make - studies of HCI cannot ignore the *Wildcard effect*. Normal, random variability among participants can result in highly variable outcomes, making the comparison of techniques difficult.

Unfortunately, the authors failed to acknowledge the limits that their highly variable results placed on statistical conclusion validity. Several comparisons were made despite unreliable estimates of group performance. Additionally, the authors appeared to recognize problems with selection (both general and specific) but failed to qualify their statements in the conclusion section. Consequently, if time-pressed practitioners were to skip the body of this study and simply read the conclusions, they would be misled.

We find the conclusion that analytic-UEMs are best used to make relative rather than absolute comparisons to be convincing. However, we disagree with the authors regarding other claims. First, we believe that the Warm group must be ignored for reasons discussed above. Second, the claim that there are differences in the performance of the Cold versus the Hot heuristic estimation groups is not well supported - especially if one were to consider relative estimates rather than absolute estimates. Third, with the key factors of training and expertise stacked against them, the KLM group did amazingly well.

5.6 Summary of the Reviews

Our disheartening conclusion is that each of these five influential studies had problems in demonstrating cause-and-effect and generality; the raison d'être for using the experimental method (see Section 3). All suffered from important problems with statistical conclusion validity. Low power (too few participants), failure to control for the Wildcard effect (by using the appropriate statistical tests), and/or the tendency to yield to the allure of reporting numerous "particularly attractive differences" afflicted each of them to varying degrees. However, even if we assume that the differences reported were real, our confidence in the internal validity of these studies is low. With the exception of Karat et al., problems with setting, instrumentation, or selection provide explanations of the results that rival those proposed by the researchers.

Conclusions regarding the generality of the findings are not much brighter. On the positive side, we find little to fault in how individual[14] experimenters handled causal construct validity and one case, Karat et al., where the issues of mono-operation and mono-method bias were directly addressed. In contrast, effect construct validity was much more problematic; it is not clear that what was being compared across UEMs was their ability to assess usability (we will have more to say about this in the next section). However, if, despite all of the above, we could accept the reported results at their face value, we would not be able to generalize beyond the specific persons, settings, and times used by the experimenters. The exception, again, is Karat et al., where we would be willing to generalize to other teams composed of a mixture of UISs, SWEs, and users.

To varying degrees, the researchers engaged in the practice of offering experience-based advice in their summary and conclusions sections. Unfortunately, they did not take the care needed to distinguish such experience-based advice from experiment-based claims.

6. OBSERVATIONS & RECOMMENDATIONS

Below we first discuss the most important issue facing usability researchers and practitioners alike; the construct of usability itself. We then take a final look at the four types of validity and the tendency to draw conclusions that are either beyond the scope of the study or contradicted by the data. For each problem type we offer recommendations for avoiding the problem and provide examples drawn from lesser cited UEM studies where such problems have been avoided.

6.1 Predicting and Measuring Usability: Effect Construct Validity

Analytic-UEMs examine the intrinsic features of an interface in an attempt to identify those that will affect usability (the pay-off) in some way; errors, speed of use, difficulty of learning, etc. The desired mapping for analytic-UEMs is from features-to-problems. Empirical-UEMs typically begin with pay-off measures and attempt to relate these measures to intrinsic features of the interface that can be changed to eliminate the pay-off problem. The desired mapping for empirical-UEMs is from problems-to-features. Although it has been shown that performance (pay-offs) can be linked to design (or intrinsic) features (see, e.g., Franzke, 1995; Gray et al., 1993), this correspondence cannot be assumed and the links must be carefully forged.

Hits, false alarms, misses, and correct rejections

It is a sure bet that no UEM is perfect; any UEM will detect some problems while missing others. Figure 6 shows a detection table for a hypothetical UEM. If the UEM claims that *A* and *B* are problems and they are: these are hits. If it claims that they are problems, and they are not: these are false alarms. Likewise, if it claims that *C* and *D* are not problems, but in truth they are; these are misses. Finally, if it claims they are not problems, and they are not; these are correct rejections.

INSERT FIGURE 6 ABOUT HERE

All four cells are important. When our UEM claims that something is a problem, how confident are we that this claim is a hit rather than a false alarm? Are we confident enough to recommend (or insist) that resources be devoted to fixing this problem? Likewise, when our UEM says that *C* and *D* are not problems, but, for example, someone else in the team (e.g., boss, marketer, SWE, etc.) thinks that they are, how confident are we in saying these are not problems? The issue facing the practitioner under these circumstances is far from academic.

Unfortunately, the problem counting approach (used in all of the above studies except Nielsen and Phillips) conflates the naming of potential problems with success. This conflation is equivalent to summing hits plus false alarms (the middle row in Figure 6), while ignoring misses and correct rejections (the bottom row of Figure 6). For a UEM to be considered reliable and valid, we need estimates of how it would fill-in all cells in Figure 6. This will not be an easy task.

Figure 6 is misleading in that it implies that we have access to *truth* as the final arbiter of usability problems. Reality is more muddled. Ideally, different UEMs would converge in identifying the same set of problems, in reality the problem sets identified by two different UEMs are only partially overlapping. If UEM-A identifies a problem not found by UEM-B, does this represent a false alarm for UEM-A or a miss for UEM-B? For example, Nielsen has stated that, “heuristic evaluation picks up minor usability problems that are often not even *seen* in actual user testing” (p. 378, 1992). In his 1994 chapter (Nielsen (1992; 1994b), he claims that “seventeen of the 40 core usability problems that had been found by heuristic evaluation were confirmed by user test” (p. 45). He argues that problems not found were not false alarms for Heuristic Evaluation but were due to the characteristics of the users who were involved in user testing. “It would therefore be impossible to find these usability problems by user testing with these users, but they are still usability problems” (p. 46).

We have some sympathy for Nielsen’s argument, since we believe that empirical-UEMs (i.e., most types of user testing) have mapping problems of their own. However, the only research that we are aware of that has attempted to link specific intrinsic features as identified by Heuristic Evaluation to specific pay-offs was done by R. W. Bailey and associates. It yielded negative conclusions.

In two independent experiments, R. W. Bailey and associates (R. W. Bailey, Allan & Raiello, 1992) performed usability tests on variations of the MANTEL system that was originally devised by Molich and Nielsen (1990). They conducted user tests for five interfaces. Five users participated in each of the tests, making five groups of five users. The first group used the original MANTEL-prime system (as built from the published specifications); the fifth group used the system as redesigned by Molich and Nielsen to fix all 29 problems identified by Heuristic Evaluation (again built from published specifications) (MANTEL-ideal).

Based upon their observations of the MANTEL-prime group, the experimenters identified and fixed two usability problems to create MANTEL-2. Group two used MANTEL-2 and observations of this group were used to identify and fix two more problems, creating MANTEL-3. Group three used MANTEL-3, resulting in the identification and fix of one more problem. MANTEL-4 was used by group 4.

Experiment 2 was similar to experiment 1. A GUI MANTEL-prime was built and subjected to Heuristic Evaluation. The 43 problems named by Heuristic Evaluation were used to build MANTEL-ideal. MANTEL-2 incorporated two changes suggested by user testing of MANTEL-prime. MANTEL-3 improved upon MANTEL-2 by fixing two problems identified by user testing and, similarly, MANTEL-4 fixed two problems identified by MANTEL-3. Both studies found an improvement from MANTEL-prime to MANTEL-2 and no improvement between MANTEL-2 and any other prototype, including MANTEL-ideal. (A strength of this report is that experiment 2 essentially replicates the experiment 1 findings using a different style of interface. This replication greatly increases the generality and construct validity of the findings.) Our biggest concern with R. W. Bailey, et al.'s study is their use of total task time as the sole measure of performance. There may well be usability problems that Heuristic Evaluation is picking up but which are not reflected by such a gross measure as total task time. However, this study stands alone as an empirical attempt to validate the recommendations of Heuristic Evaluation. As such, its predominately negative conclusions suggests that Heuristic Evaluation may name many more false alarms than hits.

Tokens, types, and categories of usability problems

The issue of mapping intrinsic features to pay-offs and pay-offs to intrinsic features is not the only one that threatens the validity of the usability effect construct. Another issue arises from the attempt to classify large numbers of individual problem tokens (e.g., an error caused by a user selecting the wrong menu item) into a small number of problem categories or types (e.g., "Be consistent").

To the naive observer it might seem obvious that the field of HCI would have a set of common categories with which to discuss one of its most basic concepts: usability. We do not. Instead we have a hodgepodge collection of *do-it-yourself* categories and various collections of *rules-of-thumb*. Our survey of recent UEM-comparison studies reveals three types of problem categories; those created in the course of the study by the researchers to account for the data they had collected (Jeffries et al., 1991; Karat et al., 1992; Smilowitz, Darnell & Benson, 1993; Virzi, 1992); established lists of guidelines or heuristics that exist in the open literature (Desurvire & Thomas, 1993; Desurvire et al., 1992; Nielsen, 1992); and one based upon theory (Cuomo & Bowen, 1994).

Developing a common categorization scheme, preferably one grounded in theory, would allow us to compare types of usability problems across different types of software and interfaces. However, although such categories may be a boon to the researcher, they may be of limited utility to the practitioner. For example, John and Mashyna (1997) argue that attempts to categorize usability problems have lead us to overestimate the success of UEMs. In a carefully analyzed case study, John compared the problems found by a Cognitive Walkthrough with those identified by user testing. Cognitive Walkthrough found 18 problems that could have been found by user testing. In contrast, user testing found 37 problems that could have been found by Cognitive Walkthrough. Out of this set of potential problems only two were the same. There were 16 problems identified by Cognitive Walkthrough that were not observed in user testing and 35 problems observed in user testing that were not predicted by Cognitive Walkthrough. John argues that the practice of reducing problem tokens to problem types or categories may mislead us into believing that different UEMs have a higher level of agreement than they actually do. She further argues that knowing what problem types an interface has is not really useful for developers. Developers need to know the specific problem (e.g., a problem with an item in a particular menu) and not the general one (e.g., “there are menu problems” or “speak the users’ language”). John’s arguments highlight yet another threat to the construct validity of common measures of usability.

Convergent measures

Attempts to derive a clear and crisp definition of usability can be aptly compared to attempts to nail a blob of Jell-O to the wall. Rather than attempting to find the one best measure, we advocate approaches that attempt to get at usability by multiple converging measures. Of the research with which we are familiar, two efforts stand out. The first, by G. Bailey (1992; 1993), used six dependent variables to measure the overall usability of various interface designs. The second, by Virzi and associates (Virzi, Sorce & Herbert, 1993), compared Heuristic Evaluation done by double experts[15], think-aloud user testing, and performance-based user testing in an attempt to identify individual usability problems.

The techniques used in Virzi's performance-based condition were those advocated by Landauer (Landauer, 1988). Users were asked to complete each task as quickly and accurately as they could, without talking aloud. The actual times taken by these participants were compared with ideal times (the time needed to complete tasks without error). To identify problems the researchers "looked for tasks and subtask times with high variability and for those that took longer than the 'ideal listener' times." As shown by Franzke (1994; 1995) this technique seems especially well suited to transform time from a rough, overall measure of performance, to a tool that focuses attention on the most problematic aspects of an interface. Such micro-analyses of payoff, when used in conjunction with analytic-UEMs, should facilitate the attempt to map problems-to-features as well as features-to-problems.

Effect construct validity-Recommendations

The issues surrounding the construct validity of effect are vitally important to the success of the UEM enterprise. Elucidating these issues should be a top priority of HCI theorists, and deriving reliable and valid means of detecting and classifying usability problems should be a major concern of the HCI research community. Researchers concerned with the effectiveness of analytic-UEMs must seek to relate intrinsic attributes to usability pay-offs. In validating these relationships, we should seek the convergence of multiple performance measures. Since we have no easy way of knowing *truth*, these measures must be carefully and painstakingly analyzed for evidence that various analytic- and empirical-UEMs do indeed converge upon the same set of usability problems. Examples of the types of analyses that we believe will be fruitful in this endeavor are provided by Virzi et al. (1993) and Franzke (1994; 1995). Such research is not quick or easy to do. However, if the HCI research community is to provide HCI practitioners with UEMs that are reliable and valid, as well as quick and easy, then these are costs that we must be prepared to accept.

6.2 Recommendations for Addressing the Four Types of Validity

Statistical conclusion validity

The recommendations in this section are simple to state: most problems with statistical conclusion validity could be avoided by using a larger sample size or using multiple measures from each UIS (or SWE) collected over several sessions[16]. The underlying issue is how to make the most of a limited access to software developers, programmers, and human factors experts at most sites. The studies that had the most problems with statistical conclusion validity tried to test too many conditions at once; however, alternatives exist. Narrowing the scope of the study by reducing the number of UEMs tested is the quickest way to increase sample size and decrease problems with statistical conclusion validity. An example of a well-conducted study with a small sample size is provided by G. Bailey (1993). G. Bailey's focused study brought multiple measures of usability (as discussed in Section 0) to bear on examining one question: does the usability of interfaces that were designed by programmers differ from those designed by human factors specialists. To answer this question, he collected data from four UISs and four SWEs over several sessions. Each participant independently built prototypes of the same system. After each declared his/her prototype ready, it was tested with three users. The videotape of each usability test session was provided to the designer without comment and without interpretation. Each designer then redesigned and retested the prototype. This cycle continued until each designer called it quits. Between 3-5 designs were developed by each designer and tested with three new user test participants on each iteration. This study yielded comparisons that were statistically reliable and valid with only four participants per each of the two conditions (UIS or SWE).

Another solution to the problem of limited access to software developers, programmers, and human factors experts is to use a small number of participants in each experiment but to replicate that experiment with different participants and, perhaps, different software systems. This last strategy has the concomitant effect of increasing our confidence in internal validity, construct validity, and external validity. (Variations of this strategy were implemented by R. W. Bailey et al., 1992; as well as by Karat et al., 1992.)

Ideally, narrowing the scope of the study would result in more focused questions. Rather than broadly asking whether UEM-A is better than UEM-B, the question might become does UEM-A find more feedback problems in walk-up and use interfaces than UEM-B? A beneficial outcome of such a focus might be the use of multiple dependent measures (more than one way of measuring the effect construct) and the avoidance of opportunistic unplanned comparisons that capitalize on chance factors. In any event, researchers and practitioners should keep in mind that, if an effect is too unstable for statistics to show a significant difference, then it is too unstable to be relied upon as guidance when selecting UEMs.

Internal validity

With a large enough sample of participants, *selection* problems can be avoided by ensuring that participants from the same participant pool are randomly assigned to conditions. In studies in which differences among participants is the independent variable (e.g., expertise), care must be taken to ensure that all else (e.g., training on the UEM, experience with the software or system being evaluated, etc.) is equivalent.

Instrumentation problems can be avoided by treating the identification, categorization, and severity rating of usability problems with the same experimental rigor called for in other parts of the design. One way to reduce instrumentation problems is to have multiple blind raters (people other than the experimenters) categorize and rate problems. In the ideal case, the raters would not have knowledge of either the conditions or participants. Additionally, the order in which problems are rated should be randomized or carefully counterbalanced across raters. Measures of interrater reliability, such as Cohen’s Kappa, also should be computed and reported.

Problems with *setting* can be avoided simply by ensuring that all participants in each UEM condition perform the experiment under the same conditions and in the same location. If circumstances do not permit holding conditions and location constant then care must be taken to ensure that each UEM is tested equally often in each condition-location combination. For example, in an attempt to test 10 UISs on each of two UEMs, we could imagine recruiting 8 UISs from one company, 8 from another, and 4 from a third; with each UEM being tested on-site. However, since the condition-location varies between companies, to be internally valid we would want to use an equal number of UISs from each company in each UEM group.

Causal construct validity

Directly comparing different ways of using a single UEM (mono-operation bias) or the effectiveness of a UEM with different types of software (mono-method bias) is more of a concern for the field as a whole than for individual researchers. (However, see R. W. Bailey, et al., 1992 and Karat, et al., 1992 for examples of how to address these concerns in a single study.) The prime responsibility of individual researchers on these issues is to provide explicit information about the exact operations and methods used.

If only one variant of a UEM is used (e.g., Heuristic Evaluation done by a small group of SWEs), the individual researchers must be explicit about the exact procedures used. Readers must be aware that terminology changes over time. They must carefully check the study’s method section to make sure that the way in which the UEM was instantiated and applied accords with their understanding of the UEM.

If only one software package is used, the researchers should try to characterize it in terms of its *use-characteristics*. Ideally, this information would be sufficient to allow other researchers and practitioners to compare their software to that used in the report. For example, knowing that the UEM has been tested using a non-visual, auditory menu to a voice-mail system may cause the researcher or practitioner to be cautious about applying the UEM to a real-time, safety-critical, display-based, command and control system. Attempts such as Olson and Moran’s (1996) to characterize when, why, and how to use UEMs are a start. However, we need a more thorough classification of use-characteristics. We might model our efforts after Green’s (1989), who has attempted to discern the underlying cognitive dimensions on which software use may vary. The issue of the cognitive dimensions of software use is a construct validity issue and one in which HCI theorists can serve both HCI researchers and practitioners.

The problem of confounding, or the interaction of different treatments, deserves separate mention. The most highly recommended cure is to use independent groups of participants so that each group is exposed to just one treatment (or UEM). If the same participants are exposed to more than one treatment, then the standard operating procedure of experimental design is to *counterbalance*. In counterbalancing, each treatment is equally likely to occur first, second, or third and explicit statistical tests can be conducted to determine if treatment type interacts with treatment order (though see, Poulton, 1982, for a discussion of cases in which counterbalancing does not prevent interactions among treatments).

External validity

External validity is another threat that is more of a concern for the field-as-a-whole than for individual researchers. The standard solution to concerns with external validity is replication. If a finding can be replicated, either by the original experimenters or by a different set of experimenters, then its external validity is bolstered. However, until replication has increased our confidence about the range of evaluators, settings, and conditions to which the results generalize, the responsibility of individual researchers is to explicitly note the possible restrictions to the scope of their findings.

Some might argue that, to affect practice, conclusions must be strongly stated: otherwise practitioners will interpret the qualifications as doubt and ignore some very fruitful techniques. We argue that if a study leads a practitioner to make a false generalization, and if this generalization has negative consequences, then the credibility of the research enterprise (and the recommended UEM) are damaged severely.

Conclusion validity

Many of the studies reviewed showed a tendency to go beyond their data in offering advice about how to do usability studies. Indeed, this practice was stoutly defended by a reviewer of an earlier version of this paper. While we can think of no defense for introducing *contradicted conclusions* (conclusions not supported by the results of the study) into a paper, we do believe that the intentions of those who *went beyond their data* (making claims not investigated in the study) to offer advice are good. Their added advice represents attempts to share non-experimentally acquired experience and expertise with practitioners. Unfortunately, most of the researchers who have done this have not clearly and explicitly separated their experiment-based claims from their experience-based advice. A notable exception to this tendency is provided by Virzi, et al. (1993). Although they do not hestitate to offer advice, their discussion of the issues carefully separates experience-based advice from experiment-based claims. We can only reiterate our earlier statement: unless such care is taken, the advice may be understood as *research findings* rather than the *researcher opinion* that it is.

6.3 Summary of Observations & Recommendations

Studies of UEMs suffer from all four types of Cook and Campbell validity. The good news is that none of the problems we found are unique to HCI and all can be overcome, or at least mitigated, by following standard, behavioral science conventions for experimental design and analysis.

Problems with the two types of cause-effect validity (statistical conclusion validity and internal validity) can be resolved by individual researchers paying more attention to methodological and statistical concerns. Likewise, many concerns with generalization (construct validity and external validity) as well as the tendency to go beyond the data in reporting conclusions can be handled by following well-known experimental design considerations and reporting conventions.

We dwelt on problems with the construct validity of effect; that is, ways of measuring usability. The studies we reviewed emphasized the problem-count approach to usability with the goal (implicit or explicit) of providing focused feedback to software designers on specific problems that if fixed would increase usability. Unfortunately, by ignoring threats to the effect construct, the message that these studies convey is an erroneous one; namely, the UEM that names the most potential problems is the most effective. If practitioners are to use such quick and easy measures with confidence then links between interface features and performance outcomes must be carefully forged.

7. CONCLUSIONS

The multitude of empirical methodologies is a strength of the HCI research community and one of our most potent methodologies is the experimental method. With Cook and Campbell we believe that “the unique purpose of experiments is to provide stronger tests of *causal* hypotheses than is permitted by other forms of research” (p. 83). Our review has unearthed no inherent obstacle to applying the experimental method to HCI topics. We draw two broad conclusions from our review.

Our first conclusion concerns the two forms of cause-effect validity: statistical conclusion validity and internal validity. The papers we reviewed have adopted methods and statistical tests that are inadequate to demonstrate cause and effect. Some might argue that this demonstrates the difficulty of doing well-controlled research in an applied setting. Although such research might be difficult to conduct, several less influential studies (e.g., G. Bailey, 1993; R. W. Bailey et al., 1992; Smilowitz et al., 1993; Virzi et al., 1993) avoid such failings, demonstrating that it is not impossible. The scope of these less influential studies seems to be smaller than the scope of the five studies we reviewed; and, unfortunately, the broader scope may be the basis for the appeal of the latter.

It has been suggested to us that having some information is preferable to having no information even when that information is bad. This is an argument that we simply do not understand. If either of us were still practitioners, and were still responsible for convincing others what changes to make and what things to leave alone, we are convinced that we would rather rely upon our own expert judgment than to base decisions upon information that we knew to be bad.

Our second conclusion is a twofold one concerning *generality*, primarily the construct validity of effect. First, in most of the studies we reviewed, usability was treated as a monolithic, atheoretical construct. However, usability pays-off by increasing performance on one or more *outcomes of interest*. These outcomes of interest vary depending upon the people and the task. For example, the outcomes of interest for safety-critical systems are very different than those for ATMs or video games. Second, the name of the game for analytic-UEMs is to predict usability pay-offs (e.g., time on task) from an examination of intrinsic features (e.g., menu organization). However, there is not a one-to-one correspondence between the two. Indeed, correspondence cannot be assumed but is a theoretical question to be resolved by empirical methods.

An interest in the design of interfaces has been a persistent HCI topic; an interest in the design of experiments has not. In this review, we have attempted to show the importance of experimental design for the HCI community. Small problems in how an experiment was designed and conducted have been shown to have large effects on what we could legitimately conclude from its outcomes. If the outcomes of these experiments were trivial then such small problems could be safely ignored. We believe that these outcomes are important; it is important to know the relationship between intrinsic features and performance pay-offs; it is important to know about a UEM’s tendency to name hits versus false alarms, to declare that a feature does not present a problem (correct rejection) versus missing features that do present problems; and it is important to know what types of UEMs work best in evaluating which types of software systems. The experimental method is a potent vehicle that can be brought to bear to address these and other core HCI issues. However, to obtain these desired pay-offs, we must pay close attention to intrinsic features of experimental design.

8. NOTES

8.1 Acknowledgments

We give our thanks to Deborah A. Boehm-Davis, William Newman, Bonnie E. John, Robin Jeffries, Gary M. Olson, Peter G. Polson, and several anonymous reviewers for their comments on earlier versions of this paper. We likewise thank the researchers of the studies we reviewed. Everyone we contacted provided complete and timely responses to our many queries. We are heartened by the cooperation provided and believe that it bodes well for the future of HCI research and practice.

8.2 Support

The writing of this review was supported in part by a grant from ONR (#N00014-95-1-0175) as well as a Fellowship from the Krasnow Institute for Advanced Studies.

8.3 Authors present address

Wayne D. Gray, Human Factors and Applied Cognitive Program, George Mason University, msn 3f5, Fairfax, VA 22030, USA. E-Mail: gray@gmu.edu. Marilyn C. Salzman, Human Factors and Applied Cognitive Program, George Mason University, msn 3f5, Fairfax, VA 22030, USA. E-Mail: msalzman@gmu.edu

9. REFERENCE

Bailey, G. D. (1992). ***Iterative methodology and designer training in human-computer interface design.*** Unpublished doctoral dissertation, New Mexico State University, Las Cruces, New Mexico.

Bailey, G. D. (1993). Iterative methodology and designer training in human-computer interface design. ***Proceedings of the ACM INTERCHI'93 Conference on Human Factors in Computing Systems,*** 198-205. New York: ACM Press.

Bailey, R. W., Allan, R. W., & Raiello, P. (1992). Usability testing vs. heuristic evaluation: a head-to-head comparison. ***Proceedings of the Human Factors Society 36th Annual Meeting,*** 409-413. Santa Monica, CA: Human Factors Society.

Bovair, S., Kieras, D. E., & Polson, P. G. (1990). The acquisition and performance of text-editing skill: A cognitive complexity analysis. ***Human-Computer Interaction, 5***(1), 1-48.

Card, S. K., Moran, T. P., & Newell, A. (1983). ***The psychology of human-computer interaction***. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, T. D., & Campbell, D. T. (1979). ***Quasi-experimentation: Design and analysis issues for field settings***. Chicago: Rand McNally.

Cuomo, D. L., & Bowen, C. B. (1994). Understanding usability issues addressed by three user-system interface evaluation technique. ***Interacting with Computers, 6***(1), 86-108.

Desurvire, H., Lawrence, D., & Atwood, M. (1991). Empiricism versus judgement: Comparing user interface evaluation methods on a new telephone-based interface. ***SIGCHI Bulletin, 23***(4), 58-59.

Desurvire, H., & Thomas, J. C. (1993). Enhancing the performance of interface evaluators using non-empirical usability methods. ***Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting,*** 1132-1136. Santa Monica, CA: Human Factors and Ergonomics Society.

Desurvire, H. W., Kondziela, J. M., & Atwood, M. E. (1992). What is gained and lost when using evaluation methods other than empirical testing. ***Proceedings of the HCI'92 Conference on People and Computers VII,*** 89-102.

Franzke, M. (1994). ***Exploration and experienced performance with display-based systems*** (Ph.D. Dissertation ICS Tech. Rpt. 94-07): University of Colorado.

Franzke, M. (1995). Turning research into practice: Characteristics of display-based interaction. ***Proceedings of the ACM CHI'95 Conference on Human Factors in Computing Systems,,*** 421-428. New York: ACM Press.

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world performance. ***Human-Computer Interaction, 8***(3), 237-309.

Green, T. R. G. (1989). Cognitive dimensions of notations. ***Proceedings of the HCI'89 Conference on People and Computers V,*** 443-460. Cambridge: Cambridge University Press.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). User interface evaluation in the real world: A comparison of four techniques. ***Proceedings of the ACM CHI'91 Conference on Human Factors in Computing Systems,*** 119-124. New York: ACM Press.

John, B. E. (1994). Toward a deeper comparison of methods: a reaction to nielsen & phillips and new data. ***Proceedings of the ACM CHI'94 Conference on Human Factors in Computing Systems,*** 285-286. New York: ACM Press.

John, B. E., & Kieras, D. E. (1996a). The GOMS family of user interface analysis techniques: Comparison and contrast. ***ACM Transactions on Computer-Human Interaction, 3***(4), 320-351.

John, B. E., & Kieras, D. E. (1996b). Using GOMS for user interface design and evaluation: Which technique? ***ACM Transactions on Computer-Human Interaction, 3***(4), 287-319.

John, B. E., & Mashyna, M. M. (1997). Evaluating a multimedia authoring tool with Cognitive Walkthrough and think-aloud user studies. ***Journal of the American Society of Information Science, 48***(9).

Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. ***Proceedings of the ACM CHI'92 Conference on Human Factors in Computing Systems,*** 397-404. New York: ACM Press.

Keppel, G., & Saufley, W. H., Jr. (1980). ***Introduction to design and analysis***. San Francisco: W. H. Freeman and Company.

Kieras, D. (1997). A guide to GOMS model usability evaluation using NGOMSL. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), ***Handbook of Human-Computer Interaction***, (Second ed., pp. 733-766). New York: Elsevier.

Landauer, T. K. (1988). Research methods in human-computer interaction. In M. Helander (Ed.), ***The handbook of human-computer interaction***, (pp. 905-928). New York: Elsevier Press.

Lewis, C., & Polson, P. G. (1992,). ***Cognitive Walkthroughs: A method for theory-based evaluation of user interfaces.*** Paper presented at the Tutorial presented at the CHI '92 Conference on Human Factors in Computing Systems, Monterey, CA.

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. ***Human Factors, 36***(2), 368-378.

Molich, R., & Nielsen, J. (1990). Improving a human-computer dialogue. ***Communications of the ACM, 33***(3), 338-348.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. ***Proceedings of the ACM CHI'92 Conference on Human Factors in Computing Systems,*** 373-380. New York: ACM Press.

Nielsen, J. (1993). ***Usability Engineering***. Boston, MA: Academic Press ISBN 0-12-518405-0.

Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. ***Proceedings of the ACM CHI'94 Conference on Human Factors in Computing Systems,*** 152-158. New York: ACM Press.

Nielsen, J. (1994b). Heuristic evaluation. In J. Nielsen & R. L. Mack (Eds.), ***Usability inspection methods***, (pp. 25-62). New York: John Wiley & Sons, Inc.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. ***Proceedings of the ACM CHI'90 Conference on Human Factors in Computing Systems,*** 249-256. New York: ACM Press.

Nielsen, J., & Phillips, V. L. (1993). Estimating the relative usability of two interfaces: heuristic, formal, and empirical methods compared. ***Proceedings of the ACM INTERCHI'93 Conference on Human Factors in Computing Systems,*** 214-221. New York: ACM Press.

Olson, J. S., & Moran, T. P. (1996). Mapping the method muddle: Guidance in using methods for user interface design. In M. Rudisill, C. Lewis, P. G. Polson, & T. D. McKay (Eds.), ***Human-Computer interface designs: Success stories, emerging methods, and real world context***, . San Francisco: Morgan Kaufmann Publishers, Inc.

Polson, P. G., Lewis, C., Rieman, J., & Wharton, C. (1992). Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. ***International Journal of Man-Machine Studies, 36***(5), 741-773.

Poulton, E. C. (1982). Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. ***Psychological Bulletin, 91***(3), 673-690.

Scriven, M. (1977). The methodology of evaluation. In A. A. Bellack & H. M. Kliebard (Eds.), ***Curriculum and evaluation***, (pp. 334-371). Berkeley: McCutchan Publishing Corporation.

Smilowitz, E. D., Darnell, M. J., & Benson, A. E. (1993). Are we overlooking some usability testing methods? a comparison of lab, beta, and forum tests. ***Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting,*** 300-303. Santa Monica, CA: Human Factors and Ergonomics Society.

Smith, S. L., & Mosier, J. N. (1986). ***Guidelines for designing user interface software*** (ESD-TR-86-278): MITRE Corporation.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? ***Human Factors, 34***(4), 457-468.

Virzi, R. A., Sorce, J. F., & Herbert, L. B. (1993). A comparison of three usability evaluation methods: Heuristic, think-aloud, and performance testing. ***Proceedings of the Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting,*** 309-313. Santa Monica, CA: Human Factors and Ergonomics Society.

Wharton, C., Bradford, J., Jeffries, R., & Franzke, M. (1992). Applying cognitive walkthroughs to more complex user interfaces: experiences, issues, and recommendations. ***Proceedings of the ACM CHI'92 Conference on Human Factors in Computing Systems,*** 381-388. New York: ACM Press.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner’s guide. In J. Nielsen & R. L. Mack (Eds.), ***Usability Inspection Methods***, . New York: John Wiley.

10. END NOTES

[1] N.B., although our focus is on studies that compare UEMs, we have no reason to believe that the problems we find in these studies are not endemic to other areas of HCI research.

[2] *Eyeball test* refers to the practice of looking at the data (eyeballing it) and deciding by intuition that differences between two raw numbers, percentages, or means are real.

[3] The term “Wildcard” is meant to apply to variability among novices as well as among experts. (Note that some decks, or studies, may have more than one Wildcard.) Our attempt at creating a mnemonic label for this effect should not obscure the basic statistical concern with the testing of group differences against participant (subject) variability.

[4] N.B., the studies we review often define two types of evaluators. One whose primary responsibility is human factors or interface issues and another whose primary responsibility is designing or writing code. For consistency’s sake we refer to these two as user interface specialists (UISs) or software engineers (SWEs).

[5] The short lists used for Heuristic Evaluation typically have less than a dozen guidelines. In contrast, the longer lists have several dozen or more guidelines. For example, one common source of guidelines (Smith & Mosier, 1986) has several hundred.

6 In the more extended version our paper, we characterized four of our current five studies as the most influential UEM-comparison studies. An additional 4 studies were listed as less influential UEM-comparisons. The final three were reviewed as UEM-expertise studies. Whereas none of these reviewers felt inclined to add studies to our list, most of them urged us to reduce the list to a core set of studies. Hence, we kept our original four UEM-comparison studies and added what we believe is the most influential of the UEM-expertise studies (Nielsen, 1992).

	Scenario	
Guidelines	**no**	**yes**
none	expert review	expert walkthrough
short list	Heuristic Evaluation	heuristic walkthrough
long list	guidelines	guidelines walkthrough
info processing perspective	N/A	Cognitive Walkthrough

Figure 1: UEM Terminology Guide

			Type of statistics	
	number of comparisons	**participants per group**	**participant variability considered?**	**eyeball or other**
Jeffries, et al., 1991	potential prob.	4-3-3-6(UT)	no	most
Karat, et al., 1992	okay	6-6-6-6-6-6	no	most
Nielsen, 1992[1]	potential prob.	31-19-14	some	many
Desurvire, et al., 1992	potential prob.	3-3-3-3-3 each; 18(UT)	no	all
Nielsen & Phillips, 1993	potential prob.	12-10-15-19-20(UT)	variance reported	most

[1] Data summary based upon the experimental part of the paper only.

Figure 2: Summary of information discussed in STATISTICAL CONCLUSION VALIDITY

	Erroneous claims due to problems with			Claims that exceed scope of study due to	
	stat concl validity	internal validity	construct validity	external validity	conclusion validity
Jeffries, et al., 1991	X	X	X	X	X
Karat, et al., 1992	X		X	X	X
Nielsen, 1992	X	X	X		X
Desurvire, et al., 1992	X	X	X	X	X
Nielsen & Phillips, 1993	X	X	X	X	X

Figure 3: Problems by validity types across studies

	User Testing	Individual Walkthrough	Team Walkthrough	total per system
System A	6	6	6 teams of 2	18
System B	6	6	6 teams of 2	18
total per UEM	12	12	12 teams of 2	

Figure 4: Participants per condition in Karat, et al., 1992.

	HE	**CW**	**Total**
UIS	3	3	**6**
non-experts	3	3	**6**
SWE	3	3	**3***

*Only 3 SWEs were used. The same 3 participated in both the Heuristic Evaluation and Cognitive Walkthrough condition.

Figure 5: Participants per group in Desurvire et al., 1992.

UEM claims that . . .	**Truth**	
	Real Problem Exists	**No Problem Exists**
***A & B* are problems**	Hit	False Alarm
***C & D* are NOT problems**	Miss	Correct Rejection

Figure 6: Usability problem detection.

Appendix A. Validity of claims made by the studies reviewed

Claim	Validity Problems				
	stat concl validity	**internal validity**	**construct validity**	**external validity**	**conclusion validity**
1) “Overall, the [H]euristic [E]valuation technique as applied here produced the best results. It found the most problems, including more of the most serious ones, than did any other technique, and at the lowest cost” p. 123	• low power • no stat tests • # of comparisons	• selection • instrument • setting		• sample	
2) Heuristic Evaluation is dependent upon “having access to several people with the knowledge and experience necessary to apply the technique” p. 123	• low power • no stat tests				• beyond the scope
3) “Another limitation of [H]euristic [E]valuation is the large number of specific, one-time, and low-priority problems found and reported.” p. 123		• instrument	• cause • effect		
4) “Usability testing did a good job of finding serious problems . . . and was very good at finding recurring and general problems, and at avoiding low-priority problems.” p. 123		• instrument	• cause • effect		
5) User testing “was the most expensive of the four techniques. . . . despite this cost, there were many serious problems that it failed to find.” p. 123	• low power • no stat tests	• selection • instrument • setting	• cause • effect	• sample	
6) “The guidelines evaluation was the best of the four techniques at finding recurring and general problems. . . . [but] missed a large number of the most severe problems.” p. 123	• low power • no stat tests	• selection • instrument • setting	• cause • effect	• sample	
7) "The [C]ognitive [W]alkthrough technique was roughly comparable in performance to guidelines. . . . In general, the problems they (Cognitive Walkthrough group) found were less general and less recurring than those found by other techniques." p. 123-124	• low power • no stat tests	• selection • instrument • setting	• cause • effect	• sample • setting	

Figure A1: Validity of claims made by Jeffries, et al., 1991.

Claim	**Validity Problems**				
	stat concl validity	**internal validity**	**construct validity**	**external validity**	**conclusion validity**
1) “Findings regarding the relative effectiveness of empirical testing and walkthrough methods were generally replicated across the two GUI systems... the significant differences in the style and presentation of the two GUI systems in the study support the reliability of the results across these types of systems” p. 402	• low power • no stat test				
2) User testing “identified the largest number of problems, and identified a significant number of relatively severe problems that were missed by the walkthrough conditions” p. 402	• random heterogen				
3) “Walkthroughs of the type in this study are a good alternative when resources are very limited.”			• effect		• beyond the scope
4) “These methods [user testing and walkthroughs] are complementary and yield different results; they act as different types of sieves in identifying usability problems” p. 403			• effect		• contradicted
5) Jeffries, et al. (1991), Desurvire, et al. (1991), as well as the current study “provide strong support for the value of [user interface] expertise” p. 403					• beyond the scope
6) “Team walkthroughs achieved better results than individual walkthroughs in some areas” p. 403	• low power • not stat tests		• effect		
7) “All walkthrough groups favored the use of scenarios over self-guided exploration in identifying usability problems. This evidence supports the use of a set of rich scenarios developed in consultation with end users.” p. 403	• no stat test		• confounding		• beyond scope
8) “The results also demonstrate that evaluators who have relevant computer experience and represent a sample of end users and development team members can complete usability walkthroughs with relative success” p. 403				• sample	

Figure A1: Validity of claims made by Karat, et al., 1992.

Claims	Validity Problems				
	stat concl validity	**internal validity**	**construct validity**	**external validity**	**conclusion validity**
1) "Usability specialists were much better than those without usability expertise at finding usability problems by heuristic evaluation." p. 380	• no stat test	• unable to evaluate	• effect		• beyond the scope
2) "usability specialists with expertise in the specific kind of interface being evaluated [double experts] did much better than regular usability specialists without such expertise [single experts], especially with regard to certain usability problems that were unique to that kind of interface." p. 380	• no stat test	• unable to evaluate	• effect		
3) "Major usability problems have a higher probability than minor problems of being found in a heuristic evaluation, but about twice as many minor problems are found in absolute numbers." p. 380	• no stat test	• instrument	• effect		
4) "Problems with the lack of clearly marked exits are harder to find than problems violating other heuristics." p. 380	• # of comparisons • no stat test	• instrument	• effect		
5) "[U]sability problems that relate to a missing interface element are harder to find when an interface is evaluated in a paper prototype form." p. 380	• # of comparisons • no stat test	• instrument	• effect		

Figure A2: Validity of claims made by Nielsen, 1992.

Claim	Validity Problems				
	stat concl validity	**internal validity**	**construct validity**	**external validity**	**conclusion validity**
1) "Heuristic Evaluation is a better method than the Cognitive Walkthrough for predicting specific problems that actually occur in the laboratory, especially for [UIS]" pp. 98-99	• low power • no stat test • # of comparisons		• treatment interact		
2) "[UIS] in the Heuristic Evaluation condition named almost twice as many problems that caused task failure or were of minor annoyance in the laboratory, than [UIS] in the cognitive condition" p. 99	• low power • no stat test • # of comparisons		• treatment interact		
3) "Heuristic Evaluation seems to facilitate the identification of potential problems and improvements that go beyond the scope of the tasks, more so than the Cognitive Walkthrough" p. 99					• beyond the scope
3) "In the Cognitive Walkthrough evaluation, [UIS] are good at predicting time, task and prompt related problems. [SWEs] are good at naming system, time, and prompt related problems. Non-Experts are only good at finding time related problems. In Heuristic Evaluation, [UIS] are good at time, prompt, task, and system related problems. [SWEs] are good at system, time, and prompt problems, and Non-Experts are best at system and keying problems" p. 99	• low power • no stat test • # of comparisons	• instrument	• treatment interact		
4) "[UIS] focused on problems that violate the heuristic, 'provide feedback', where they are more focused on the user's interaction with the system than the [SWEs] who were . . ." p. 99	• low power • no stat test • # of comparisons				
5) "[UIS] were the best at predicting laboratory problems that caused task failure, errors, and caused confusion in the users. The [UIS] were better in the heuristic condition than in the Cognitive Walkthrough, and there were no differences between methods for the [SWEs] and the Non-Experts" p. 99	• low power • no stat test • # of comparisons	• instrument	• treatment interact		
6) "[UIS] were also best at predicting the user's attitude as a result of a problem in the laboratory." p. 99	• low power • no stat test				
7) "This study has shown that evaluation methods can identify a number of interface problems, and these methods are particularly useful by [UIS]" p. 100.	• low power • no stat test • # of comparisons		• treatment interact		
8) "At best, these methods provide only 44% of the problems seen in a laboratory based usability study" p. 100.	• low power • no stat test • # of comparisons				

Figure A3: Validity of claims made by Desurvire, et al., 1992.

Claim	Validity Problems				
	stat concl validity	**internal validity**	**construct validity**	**external validity**	**beyond the data**
1) “User testing still seems to be the best method for arriving at [estimates of user performance], but one should remember that laboratory testing is not always a perfect predictor of field performance” p. 220.	• no stat tests		• effect		
2) “User testing was also much more expensive than ‘cold’ heuristic estimates and somewhat more expensive than [KLM] analyses” pp. 220-221		• selection • setting	• effect	• represent sample	
3) “Heuristic estimates were better in the hot condition where estimators had access to running versions of the two interfaces, than in the cold condition based on specifications only” p. 221	• no stat tests				
4) “Estimates of the relative advantage of one interface over the other were much better than estimates of the absolute time needed by users to perform various tasks” p. 221.	• no stat tests				
5) “[KLM] and heuristic estimates were about equal for relative estimates, so heuristic estimates might be recommended based on its lower costs” p. 221		• selection • setting • instrument	• effect	• represent sample	• beyond the scope
6) “[KLM] analyses were superior for absolute estimates” p. 221.	• no stat tests	• selection			
7) “Performance estimates from both heuristic estimation and [KLM] analyses are highly variable” p. 221		• selection • setting		• represent sample	• contradicted

Figure A4: Validity of claims made by Nielsen and Phillips, 1993.