

CRITERIA FOR EVALUATING USABILITY EVALUATION METHODS

H. Rex Hartson[1], Terence S. Andre[2], and Robert C. Williges[2]

[1]Department of Computer Science
[2]Department of Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA 24061 USA
hartson@vt.edu, tandre@vt.edu, williges@vt.edu

ABSTRACT

Among interactive system developers and users there is now much agreement that usability is an essential quality of software systems. The introduction of usability evaluation methods (UEMs) to assess and improve usability in such systems has led to a variety of alternative approaches and a general lack of understanding of the capabilities and limitations of each. This confusion has intensified the need for practitioners and others to determine which methods are more effective, and in what ways and for what purposes. However, UEMs cannot be evaluated and compared reliably because of the lack of standard criteria for comparison. This paper presents a practical discussion of factors, comparison criteria, and UEM performance measures that are interesting and useful in studies comparing UEMs. We have attempted to highlight major considerations and concepts, offering some operational definitions and exposing the hazards of some approaches proposed or reported in the literature. In demonstrating the importance of developing appropriate UEM evaluation criteria, we present some different possible measures of effectiveness, select and review studies that use two of the more popular measures, and consider the trade-offs among different criterion definitions. In sum, this work highlights some of the specific challenges that researchers and practitioners face when comparing UEMs and provides a point of departure for further discussion and refinement of the principles and techniques used to approach UEM evaluation and comparison.

1. INTRODUCTION

The concept of evaluation dates back to the beginning of system analysis and human factors and beyond. Usability evaluation reaches back to virtually the beginning of human-computer interaction (HCI), usability evaluation methods go back more than a decade (Card, Moran, & Newell, 1983; Nielsen & Molich, 1990), and studies for comparing UEMs have been conducted for some time, too (Nielsen & Molich, 1990; Jeffries, Miller, Wharton, & Uyeda, 1991). However, in a broad historical view, the area is still relatively new and incomplete as both a research topic and as an applied body of knowledge.

In the 1980s, laboratory usability testing quickly became the primary UEM for examining a new or modified interface. Laboratory usability testing was seen by developers as a way to minimize the cost of service calls, increase sales through the design of a more competitive product, minimize risk, and create a historical record of usability benchmarks for future releases (Rubin, 1994). Laboratory testing involved user performance testing to evaluate speed, accuracy, and errors in addition to user subjective evaluations. User-based evaluation methods included verbal protocols (Ericsson & Simon, 1984), critical incident reporting (del Galdo, Williges, Williges, & Wixon,

1987), and user satisfaction ratings (Chin, Diehl, & Norman, 1988). In the 1990s, many developers have explored other methods in an attempt to bring down the cost and time requirements of traditional usability testing. In addition, because usability testing often occurs late in the design process, developers were motivated to look at methods that could be used earlier when only an immature design was available (Marchetti, 1994). As a result, expert-based inspection methods grew in popularity because many of them were intended to be used with a relatively early design concept (Bradford, 1994). Some of the more popular expert-based UEMs include: Guideline Reviews based on interaction design guidelines such as those by Smith and Mosier (1986), Heuristic Evaluation (Nielsen & Molich, 1990), Cognitive Walkthroughs (Lewis, Polson, Wharton, & Rieman, 1990; Wharton, Bradford, Jeffries, & Franzke, 1992) Usability Walkthroughs (Bias, 1991), Formal Usability Inspections (Kahn & Prail, 1994), and Heuristic Walkthroughs (Sears, 1997).

Practitioners are far from settled on a uniform UEM and researchers are far from agreement on a standard means for evaluating and comparing UEMs. Confounding the situation is a miscomprehension of the limitations of UEMs and under what conditions those limitations apply. What makes the state of UEM affairs most disturbing to many is the lack of understanding of UEM evaluation and comparison studies, as pointed out by Gray and Salzman (1998) and debated in subsequent discussion (Olson & Moran, 1998). Gray and Salzman call into question existing UEM studies as "potentially misleading" and urge researchers to apply the power of the experimental method carefully and rigorously in these studies.

1.1. Iterative Design and Evaluation

Interactive systems are usually designed through an iterative process involving design, evaluation, and redesign. Kies, Williges, and Rosson (1998) summarized three major iterative stages of initial, prototype, and final design that are central to the iterative design process. During initial design, goals and guidelines are iterated to finalize the design specifications leading to a prototype design. Formative evaluation focuses on usability problems that need to be solved during the prototype design stage before a final design can be accepted for release. Summative evaluation is then conducted to evaluate the efficacy of the final design or to compare competing design alternatives in terms of usability. As shown in Figure 1, UEMs are used primarily for formative evaluations during the prototype design stage. These formative evaluations are focused on efficient and effective techniques to determine usability problems that need to be eliminated through redesign. A combination of expert-based and user-based inspection methods has evolved to facilitate the formative evaluation process.

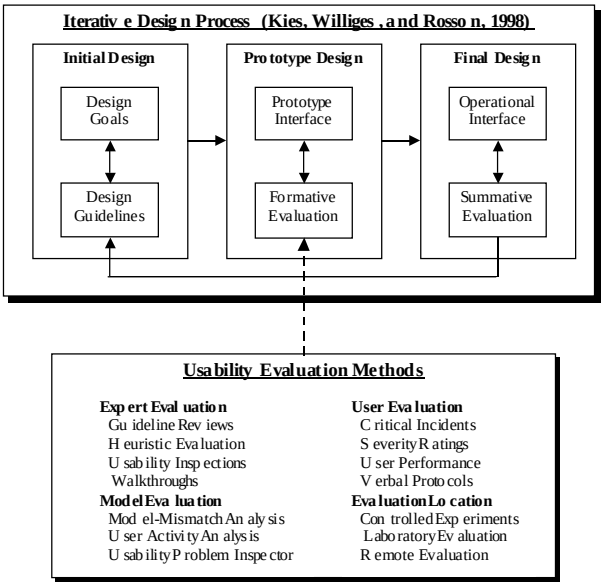

Figure 1. UEMs Used in Formative Usability Evaluation.

1.2. The Need for a Foundation for Evaluating Usability Evaluation Methods

Among interactive system developers and users there is now much agreement that usability is an essential quality of software systems. Among the HCI and usability communities, there is also much agreement that:

- usability is seated in the interaction design,
- an iterative, evaluation-centered process is essential for developing high usability in interaction designs,
- usability, or at least usability indicators, can be viewed as quantitative and measurable, and
- a class of usability techniques called UEMs have emerged to carry out essential usability evaluation and measurement activities.

Beyond this level of agreement, however, there are many ways to evaluate the usability of an interaction design (i.e., many UEMs), and there is much room for disagreement and discussion about the relative merits of the various UEMs. As more new methods are being introduced, the variety of alternative approaches and a general lack of understanding of the capabilities and limitations of each has intensified the need for practitioners and others to determine which methods are more effective, and in what ways and for what purposes. In reality, researchers find it difficult to reliably compare UEMs because of a lack of:

- standard criteria for comparison,
- standard definitions, measures, and metrics on which to base the criteria, and
- stable, standard processes for UEM evaluation and comparison.

Lund (1998) noted the need for a standardized set of usability metrics, citing the difficulty in comparing various UEMs and measures of usability effectiveness. As Lund points out, there is no single standard for direct comparison, resulting in a multiplicity of different measures used in the studies, capturing different data defined in different ways. Consequently very few studies clearly

identify the target criteria against which to measure success of a UEM being examined. As a result, the body of literature reporting UEM comparison studies does not support accurate or meaningful assessment or comparisons among UEMs.

Some have tried to help by performing UEM comparison studies, but many such studies that have been reported were not complete or otherwise fell short of the kind of scientific contribution needed. Although these shortcomings often stemmed from practical constraints, they have led to substantial critical discussion in the HCI literature (Gray & Salzman, 1998; Olson & Moran, 1998).

Accordingly, this paper presents a practical discussion of factors, comparison criteria, and UEM performance measures that are interesting and useful in studies comparing UEMs. We have attempted to highlight major considerations and concepts, offering some operational definitions and exposing the hazards of some approaches proposed or reported in the literature. In demonstrating the importance of developing appropriate UEM evaluation criteria, we present some different possible measures of effectiveness, select and review studies that use two of the more popular measures, and consider the trade-offs among different criterion definitions. This work highlights some of the specific challenges that researchers and practitioners face when comparing UEMs and provides a point of departure for further discussion and refinement of the principles and techniques used to approach UEM evaluation and comparison.

1.3. Terminology

In the context of this paper, the term *usability evaluation method* (UEM) is taken to refer to any method or technique used to perform usability evaluation, with emphasis on formative usability evaluation (i.e., usability evaluation/testing used to improve usability) of an interaction design at any stage of its development. This broad definition includes lab-based usability testing with users, heuristic and other expert-based usability inspection methods, model-based analytic methods, all kinds of expert evaluation, and remote evaluation of interactive software after deployment in the field. The essential common characteristic, for purposes of this paper, is that every method, when applied to an interaction design, produces a list of potential usability problems as its output. Some UEMs have additional functionality, such as the ability to help write usability problem reports, to classify usability problems by type, to map problems to causative features in the design, or to offer redesign suggestions. We believe these are all important and deserve attention in addition to the basic performance-based studies.

A person using a UEM to evaluate usability of an interaction design is called an *evaluator*, to distinguish this specialized usability engineering role. More specifically, a person using a usability inspection method (one type of UEM) is often called an *inspector*.

In this work we use the term “UEM comparison study” to refer to any empirical summative evaluation that compares performance (by any measure) among UEMs. Statistical significance, while not an absolute requirement, is a strong expectation.

1.4. Types of Evaluation and Types of UEMs

In order to understand UEMs and their evaluation, one must understand evaluation in the context of usability. We have adopted Scriven's (1967) distinction between two basic approaches to evaluation based on the evaluation objective. *Formative evaluation is evaluation done during development to improve a design and summative evaluation is evaluation done after development to assess a design (absolute or comparative).* Phrasing Scriven's definitions in terms of usability,

formative evaluation is used to find usability problems to fix so that an interaction design can be improved. Summative evaluation is used to assess and/or compare the level of usability achieved in an interaction design. UEMs are used to perform formative, not summative, usability evaluation of interaction designs. Formal experimental design, including a test for statistical significance, is used to perform summative evaluation and is often used to compare design factors in a way that can add to the accumulated knowledge within the field of HCI.

Sometimes formative usability evaluation can also have a component with a summative flavor. Some UEMs support collection of quantitative usability data in addition to the qualitative data (e.g., usability problem lists). For example, measurement of user task performance quantitatively in terms of time-on-task and error rates adds a summative flavor to the formative process because it is used to assess the level of usability. Not being statistically significant, these results do not contribute (directly) to the science of usability, but are valuable usability *engineering* measures within a development project. Usability engineers, managers, and marketing people use quantitative usability data to identify convergence of a design to an acceptable level of usability, to know when to stop iterating the development process, and to attain a competitive edge in marketing a product.

A somewhat orthogonal perspective is used to distinguish evaluation methods in terms of how evaluation is done. Hix and Hartson (1993) describe two kinds of evaluation: analytic and empirical. Analytic evaluation is based on analysis of the characteristics of a design, through examination of a design representation, prototype, or implementation. Empirical evaluation is based on observation of the performance of the design in use. Perhaps Scriven (1967), as described by Carroll, Singley, and Rosson (1992), gets at the essence of the differences better by calling these types of evaluation, respectively, intrinsic evaluation and pay-off evaluation. Intrinsic evaluation is accomplished by way of an examination and analysis of the attributes of a design without actually putting the design to work, whereas pay-off evaluation is evaluation situated in observed usage.

Sidebar on the importance of articulating usability goals to establish appropriate criteria

In describing Scriven's distinction between intrinsic and pay-off approaches to evaluation, other authors (for example, (Carroll, et al., 1992) and (Gray & Salzman, 1998)), quote his example featuring an axe:

> 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.

While this example serves Scriven's purpose well, it also offers us a chance to make a point about the need to carefully identify usability goals before establishing evaluation criteria. Giving an HCI usability perspective to the axe example, we see that pay-off evaluation does not necessitate observation of axe performance *in the hands of an expert axeman* (or axeperson). Expert usage might be one component of the vision in axe design for usability, but is not an essential part of the definition of pay-off evaluation. Usability goals depend on expected user classes and the expected kind of usage. For example, an axe design that gives optimum performance in the hands of an expert might be too dangerous for a novice user. For the city dweller, a.k.a. weekend wood whacker, safety might be a usability goal that transcends firewood production, calling for a safer design that might necessarily sacrifice efficiency. One hesitates to contemplate the metric for this case, possibly counting the number of 911 calls from a cell phone in the woods. Analogously in the user interface

domain, usability goals for a novice user of a software accounting system, for example, might place data integrity (error avoidance) higher in importance than sheer productivity.

The de facto standard pay-off method in the usability world is the well-known lab-based usability testing with users. GOMS (Card, et al., 1983) analysis, where user actions for task performance are assigned costs (in terms of time), set within a model of the human as information processor, offers a good representative example of intrinsic usability evaluation. Some usability inspection methods (Nielsen & Mack, 1994) are essentially intrinsic in that they analyze an interaction design with respect to a set of design guidelines or "heuristics." This kind of inspection method requires a usability expert to analyze the design rather than testing with real users. Other usability inspections methods are hybrids between intrinsic and pay-off in that the analysis done during usability inspection is task driven; the expert's analysis is based on exploring task performance and encountering usability problems in much the same way users would, adding a pay-off dimension to the intrinsic analysis. In this situation, a usability inspector asks questions about designs in the context of tasks in order to predict problems users would have.

Regardless of the method, the goal of all UEMs is essentially the same: to produce descriptions of usability problems observed or detected in the interaction design, for the purpose of analysis and redesign. Ostensibly, this shared goal and common output should make the various UEMs directly comparable but, as we already know from the discussion in the literature, things are not that simple.

1.5. Damaged Goods

The need to evaluate and compare UEMs is underscored by the fact that some developers have recently questioned the effectiveness of some types of UEMs in terms of their ability to predict problems that users actually encounter (John & Marks, 1997). Gray and Salzman (1998) recently documented, in their article about "damaged goods," specific validity concerns about five popular UEM comparison studies. A key concern noted by Gray and Salzman is the issue of using the right measure (or measures) to compare UEMs in terms of effectiveness.

To be fair, some of the incomplete results criticized by Gray and Salzman were understandable because researchers were using data that became available through means designed for other ends (e.g., usability testing within a real development project), and additional resources were not available to conduct a complete, scientifically valid experiment. It is fair to say that these partial results have value as indicators of relative UEM merit in a field where complete scientific results are scarce indeed. Nonetheless, there is certainly a need for more carefully designed comparison experiments – both to contribute to the science of usability and to provide practitioners with more reliable information about the relative performance of various UEMs as used for various purposes.

One could apply the term "damaged goods" to results generated by UEMs themselves, since their application represents a kind of empirical "study." In the perspective of a trained human factors engineer, a study not rigorously designed and executed according to the prescripts of experimental design methodology for statistical significance, is damaged goods. However, in many engineering contexts, usability engineering included, damaged goods are not always a bad thing. Most UEMs represent a conscious tradeoff of performance for savings in cost. As long as "buyers" know what they are getting and it will suffice for their needs, they can often get a good "price" for damaged goods. This is, we think, a sound principle behind what has been called *discount engineering*

(Nielsen, 1989), and has always been part of the legitimate difference between science and engineering.

However, as pointed out by Gray and Salzman, damaged goods are far less acceptable in the realm of usability science, especially when found in the form of poorly designed UEM comparison studies. Some authors in the sequel discussion of the Gray and Salzman article (Olson & Moran, 1998) suggest that some science is better than none and resource limitations that preclude complete scientific results should not prevent attempts at modest contributions. These discussants argue that this is especially true in a relatively new field where any kind of results is difficult to come by. In balance, Gray and Salzman would probably caution us that sometimes bad science is worse than none. But, as Lund (1998) points out in his commentary about Gray and Salzman, the danger may not loom so darkly to practitioners, making the case that practitioners will quickly discover if a recommendation is not useful.

In any case, the argument made above, which applies to the acceptance of any goods, was based on the "buyers" knowing what they are, and are not, getting for their money. That means researchers must develop effective UEM evaluation criteria and it means practitioners must understand those criteria and how they apply.

1.6. Roadmap of Concepts

Figure 2 shows a guide to the concepts of this paper and the relationships among them. Researchers planning a UEM comparison study have in mind an ultimate criterion for establishing the "goodness" of a particular method. However, since the ultimate criterion is generally considered theoretically impossible to measure, researchers select one of many possible actual criteria to approximate the ultimate criterion for UEM comparison. The experimenter then applies a method representing the actual criterion to identify a "standard" set of usability problems existing in the target system interaction design. The experimenter also applies the UEMs being compared (UEM_A and UEM_B in Figure 2) to the target design and calculates UEM performance metrics using the resulting usability problem lists in relation to the "standard" usability problem list. The UEMs are then compared on the basis of their performance metrics. These concepts are all discussed in detail in the sections that follow.

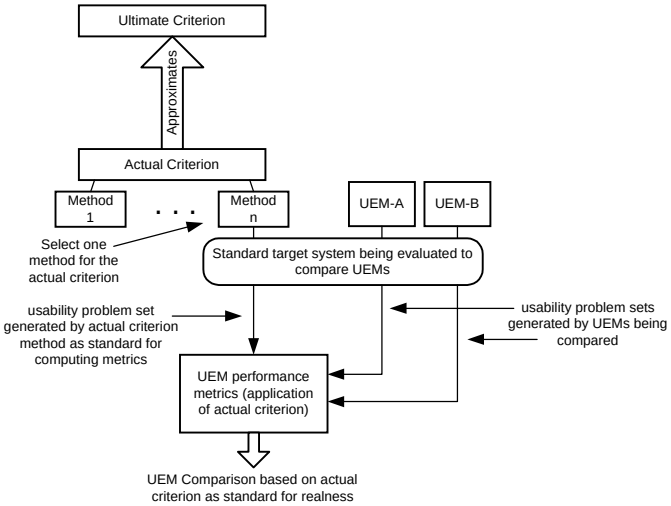

Figure 2. Roadmap of Concepts.

2. THE EVALUATION CRITERION PROBLEM

To evaluate the effectiveness of a UEM, and especially to compare the effectiveness of UEMs, usability researchers must establish a definition for effectiveness and an evaluation or comparison criterion, or criteria. The criteria are stated in terms of one or more performance-related (UEM performance, not user performance) measures (effectiveness indicators), that are computed from raw empirical usability data (e.g., usability problem lists) yielded by each UEM. Making the right choice for criteria and performance measures depends on understanding the alternatives available and the limitations of each. In this paper, we bring these issues to light to foster this understanding.

2.1. Criterion Relevance, Deficiency, and Contamination

The selection of criteria to evaluate a UEM is not essentially different from criteria selection for evaluation of other kinds of systems (Meister, Andre, & Aretz, 1997). In the evaluation of large-scale systems such as military weapon systems, for example, customers (e.g., the military commanders) establish *ultimate* criteria for a system in the real world. Ultimate criteria are usually simple and direct – for example, that a certain weapon system will win a battle under specified conditions. However, military commanders cannot measure such ultimate criteria directly outside of an actual combat environment. As a result, military commanders establish specific other attributes, called *actual* criteria, which are more easily measured and which there is reason to believe will be effective predictors of the ultimate criteria. To illustrate, commanders might establish the following characteristics as actual criteria for military aircraft performance: aircraft must fly at X thousand feet, move at Y mach speed, and shoot with Z accuracy. As actual criteria, these measures are only indicators or predictors of the ultimate criterion and are more valuable as predictors if they can be validated, which can happen only when real combat clashes occur.

Measures to be used in actual criteria are designed to be operational parameters that can be computed by consistent means that are agreed upon and reliably understood. If system reliability was a goal, for example, mean-time-between-failure (MTBF), is a good measure because practitioners understand its meaning and computational mechanics. Researchers do not have any measures this well standardized yet in usability, so they generally define their own to meet the goals of the research (Gray & Salzman, 1998; Lund, 1998). To be useful and repeatable in an actual criterion, a measure must have at least these characteristics:

- a solid definition, understandable by all,
- a metric, to be computed from raw usability data,
- a standard way to measure or take data, and
- one or more levels of performance that can be taken as a "score" to indicate "goodness."

The degree to which actual criteria are successful predictors of the ultimate criterion is the essence of the concept called *criterion relevance*, illustrated by the intersection of the two circles in Figure 3. If, for example, stealth technology makes it unnecessary to fly at 80,000 feet, then the altitude criterion is no longer a useful predictor of the ultimate criterion causing that part of the actual criterion to fall outside the intersection with the ultimate criterion. Because this part of the actual criterion contaminates the approximation to the ultimate criterion, it is called *criterion contamination*.

If military commanders leave out an important measure that should be included in the estimate of ultimate criterion, the actual criterion is deficient in representing the ultimate criterion and the part of the ultimate criterion not represented falls outside the intersection in the part called *criterion deficiency*.

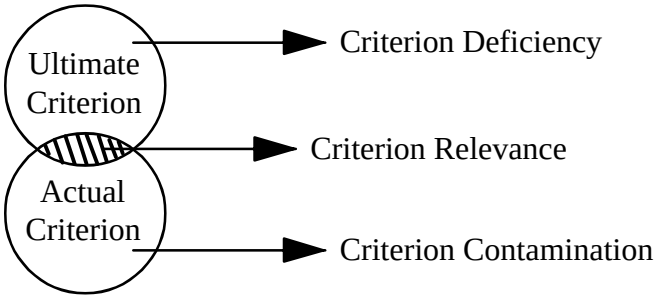

Figure 3. Relationship Between Ultimate and Actual Criteria.

3. ULTIMATE CRITERION FOR UEM EFFECTIVENESS – FINDING REAL USABILITY PROBLEMS

Criterion relevance applies to UEMs as well as military planes. For the purpose of discussion, we postulate the following ultimate criterion for UEM evaluation and comparison, somewhat analogous to the simple ultimate criterion used in the case of the airplane: *How well does the UEM help inspectors or evaluators discover real usability problems?*

This realness attribute, which plays a pivotal role in several of the UEM measures is defined as follows: *A usability problem (e.g., found by a UEM) is real if it is a predictor of a problem that users will encounter in real work-context usage and that will have an impact on usability (user performance, productivity, and/or satisfaction).* This would exclude problems with trivially low impact and situations real users would/could not encounter. The emphasis on real users is important in this definition, because many of the UEMs evaluated in studies are usability inspection methods, where the inspectors encounter problems that do not always predict usability problems for real users. In any case, this definition of realness belongs more to ultimate criteria than to any actual criterion, since it does not yet offer an operational way to test for the stated conditions. This deceptively simple UEM criterion translates into a large number of issues when it comes to putting it into practice, when it comes to unpacking the meaning of the words “how well” and “real.”

To the extent that any practical means for determining realness in the actual criterion will result in some errors, there will be both criterion contamination and criterion deficiencies. But once the actual criterion (including the method for determining realness) is established, those issues essentially disappear from immediate consideration and the focus is on the actual criterion as the “standard:” How well does the UEM help inspectors discover “real” (as determined by the method of the actual criterion) usability problems?

4. ACTUAL CRITERIA FOR UEM EFFECTIVENESS – OPERATIONALLY DETERMINING REALNESS

4.1. Determining Realness by Comparing with a Standard Usability Problem List

If an evaluator had a complete list of all the real usability problems that exist in a given target interaction design, that evaluator could ascertain the realness of each candidate usability problem found by a UEM. The evaluator would search the standard list for a match to the candidate problem, thereby determining whether it is in the list (and, thus, whether it is real).

4.1.1. Usability Problem Lists as Usability Problem Sets

In practice, each UEM produces a list of usability problems. Comparison of UEMs requires comparison and manipulation of their usability problem lists. It is useful to think of each UEM as producing a set of usability problems, because it allows us to think of the list as unordered and allows formal expressions of important questions. Cockton and Lavery (1999) favor this same choice of terminology for much the same reasons. For example, one might need to ask whether a given UEM finds a certain known problem in a target design. Or one might need to know what usability problems the outputs of UEM_1 and UEM_2 have in common, or what do you get when you merge the outputs of UEM_1 and UEM_2. These are questions about set membership, set intersections, and set unions. Thus, viewing usability problem lists as sets affords simple set operations such as union, intersection, and set difference to manipulate the usability problems and combine them in various ways to calculate UEM performance measures.

4.1.2. Producing a Standard Usability Problem Set

The second technique for determining realness requires the experimenters to establish a "standard" UEM that can be used to generate, as a comparison standard, a touchstone set of usability problems deemed to be "the real usability problems" existing in the target interaction design of the study. This standard usability problem set will be used as a basis for computing various performance measures as parts of actual criteria. We say that the touchstone set is part of an actual criterion because it can only approximate the theoretical ultimate real usability problem set, a set that cannot be computed. Some of the possible ways to produce a standard-of-comparison usability problem set for a given target interaction design include:

- seeding with known usability problems,
- laboratory-based usability testing,
- asymptotic laboratory -based testing, and
- union of usability problem sets over UEMs being compared.

The seeding approach introduces a known usability problem set to be used directly for comparison as part of the actual criterion. The two kinds of lab-testing involve users and expert observers to produce standard usability problem sets found in the target system. The union of usability problem sets combines all the problem sets produced by the UEMs to produce a standard of comparison.

4.1.2.1. Seeding the Target Design with Usability Problems

Sometimes experimenters will "seed" or "salt" a target system with known usability problems, an approach that can seem attractive because it gives control over the criterion. In fact, this is one of the few ways the experimenters can know about all the existing problems (assuming there are no real problems in the system before the seeding). But many UEM researchers believe salting the target system is not a good basis for the science of a UEM study, because the outcome depends heavily on experimenter skill (in the salting), putting ecological validity in doubt. Experienced usability practitioners will know that contrived data can seldom match the variability, surprises, and realness of usability data from a usability lab.

4.1.2.2. Laboratory-Based Usability Testing

Traditional lab-based usability testing is the de facto standard, or the "gold standard" (Landauer, 1995; Newman, 1998), used most often in studies of UEM performance. Lab-based testing is a UEM that produces high quality, but expensive, usability problem sets. Often lab-based UEM performance is unquestioned and thought of as an effective actual criterion, suitable for use as a standard of comparison to evaluate other UEMs. Because it is such a well-established comparison standard, it might be thought of as an ultimate criterion, especially when compared to usability inspection methods. However, it does not meet our definition for an ultimate criterion. In the usability lab users are constrained and controlled. Developers decide which tasks users should perform and what their work environment will be like (usually just the lab itself). Some researchers and practitioners would like more data on how well lab-based testing is predictive of real usability problems and under what conditions it best plays this role, but it is difficult to find an experimental standard good enough to make that comparison.

Despite these possible deviations from the ultimate, the experience of the usability community with lab-based testing as a mainstream UEM for formative evaluation within the interaction development process has led to a high level of confidence in this UEM. Other UEMs have arisen, not because of a search for higher quality but mostly out of a need for lower cost alternatives.

In any case, the typical lab-based usability test employs several users as subjects along with one or more observers and produces the union of problems found by all users. Given that some usability problems, even from lab-based testing, can be of questionable realness, it is best to combine the lab test with expert review to eliminate some of the problems considered not real, thus improving the quality of the usability problem set to be used as the actual criterion.

4.1.2.3. Asymptotic Lab-Based Testing

The typical usability lab test will miss some usability problems. In fact, most laboratory tests are deliberately designed with an objective of cost-effectiveness, at an acknowledged penalty of missing some usability problems. Using the formula $1 - (1 - p)^n$, researchers have shown that a sample size of five evaluators (n) is sufficient to find approximately 80% of the usability problems in a system if the average individual detection rate (p) is at least 0.30 (Virzi, 1990; Wright & Monk, 1991; Virzi, 1992; Nielsen, 1994). Virzi (1992) found average individual detection rates ranging from 0.32 to 0.42. However, Lewis (1994) found that average detection rates can be as low as 0.16 in office applications. Figure 4 shows the problem discovery likelihood when individual detection

rates range between a low of 0.15 and a high of 0.45 using the formula $1 - (1 - p)^n$. The rate at which problem detection approaches the asymptote varies significantly, depending on the individual detection rate. Only 8 evaluators are needed to find 95% of the problems when the detection rate is 0.45, while as many as 19 evaluators are needed to find the same amount when the detection rate is 0.15.

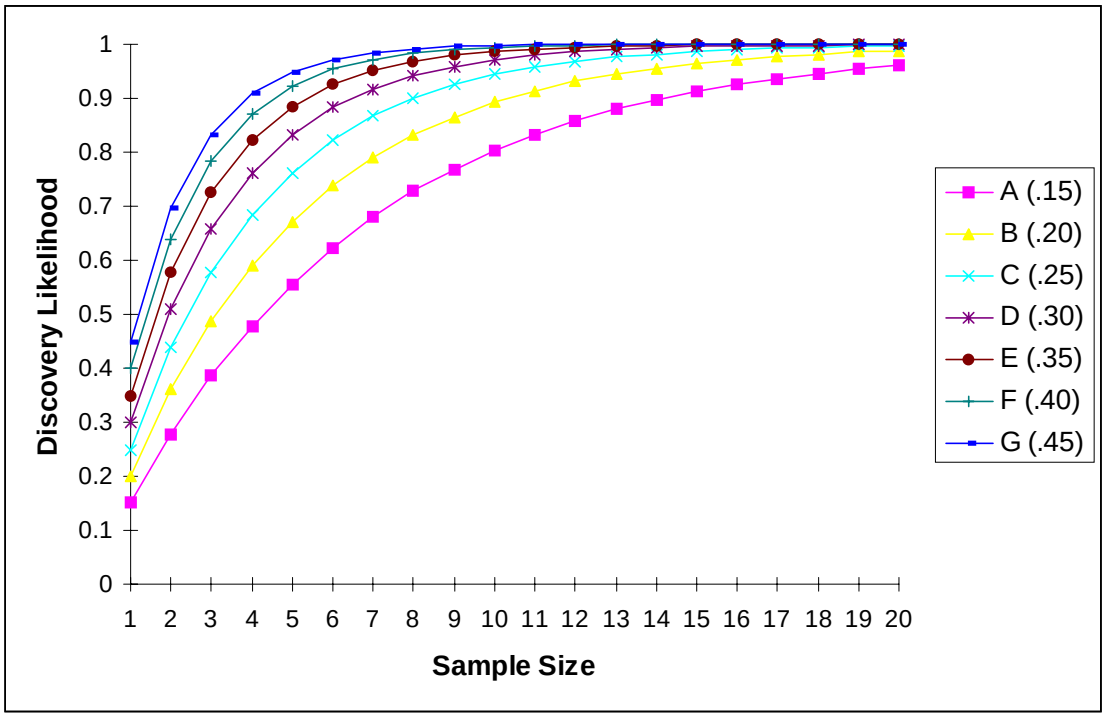

Figure 4. The Asymptotic Behavior of Discovery Likelihood as a Function of the Number of Users (adapted from Lewis, 1994).

For an individual detection rate of about 0.3 or higher, the first 3-5 users are enough to find 80% of the usability problems, as found independently by Nielsen (1990; 1992) and Virzi (1992). The number of new problems found by each added user levels off at about three-to-five users, with the number of new usability problems found dropping with each new user added after that. Therefore efficiency, or cost effectiveness, also levels off at three-to-five users. Fortunately for usability practice, both Nielsen and Virzi found that this kind of laboratory-based testing also has a tendency to find the high-severity usability problems first (Nielsen, 1992).

The total usability problems found does level off asymptotically as the number of users increases. This means that the asymptotic level can be thought of as a good approximation to the level of the ultimate criterion (after any non-real problems are removed). Thus, extending the usual laboratory-based usability test to include several more users is a good, but expensive, choice for producing a "standard" usability problem set from the target design as part of an actual criterion.

4.1.2.4. The Union of Usability Problem Sets

Another technique often used to produce a standard usability problem set as a criterion for being real is the union set of all the individual usability problem sets, as found by each of the

methods being compared (Sears, 1997). This method has a very serious drawback in that it eliminates the possibility to consider validity as a UEM measure, because the basis for metrics is not independent of the data.

4.1.3. Comparing Usability Problem Descriptions

Gray and Salzman (1998) correctly criticize just counting usability problems for UEM measures, without determining if some usability problems found overlap or duplicate others. A determination of overlap cannot be made, though, without an ability to compare usability problem descriptions. Determining realness by comparing with a standard usability problem set also requires comparison. Comparison requires complete, unambiguous usability problem descriptions that facilitate distinguishing different types of usability problems.

This comparison is straightforward in abstract sets, where each element is unambiguously identified by name or value. If $x \in A$ and $x \in B$, then the appearance of x in A is identical to its appearance in B. However, usability problem sets from UEMs are more difficult to compare because they involve enumerated sets in which elements are represented by narrative problem descriptions and elements, not having a unique canonical identity.

Because usability problem descriptions are usually written in an ad hoc manner, expressed in whatever terms seemed salient to the evaluator at the time the problem is observed, it is not unusual for two observers to write substantially different descriptions of the same problem. Thus, in order to perform set operations on usability problem sets, one needs the ability to determine when two different usability problem descriptions are referring to the same underlying usability problem. This kind of comparison of textual problem descriptions is usually done by expert judgment, but is subject to much variability. There is a need for a standard way to describe usability problems, for a framework within which usability problem descriptions can be more easily and more directly compared. We are working on just such a framework, called the User Action Framework (Hartson, Andre, Williges, & van Rens, 1999).

4.2. Determining Realness by Expert Review and Judgment

In order to determine the realness of usability problems by review and judgment of expert(s), each candidate usability problem is examined by one or more usability experts and determined by some guideline to be real or not. This technique can also have the effect of accumulating a standard list, if the judgment results can be saved and reused. This technique can also be combined with the techniques described in the following sections to filter their "standard" usability problem lists, ensuring that the results are, by their judgment, real.

Often designers of UEM studies find that the guidelines for realness to be used in expert judgment are too vague or general to be applied reliably and the judgments can vary with the expert and other experimental conditions. This introduces the possibility of a bias causing the usability problem lists of each UEM to be judged differently. As an alternative, the experimenters seek a "standard" usability problem list as a single standard against which to compare each UEM's output. This approach also involves judgment, when it comes to comparing each usability problem against the standard list.

4.3. Determining Realness by End-User Review and Judgment

Since usability is ultimately determined by the end-user not an expert evaluator, realness of problems needs to be established by the user. Specific UEM procedures have been adopted to enhance realness criteria based on problems specified by the actual user.

4.3.1. Critical Incidents

Often, verbal protocols provided by the user do not provide succinct problem descriptions. Del Galdo, Williges, Williges, & Wixon (1986) modified the critical incident technique described by Flanagan (1954) to allow users the capability of specify extremely good and poor characteristic of HCI interfaces during usability evaluation. These critical incidents can be used to classify usability problems during formative evaluation. Consequently, the critical incident becomes the actual criterion for specifying real usability problems. Subsequently, critical incidents in HCI evaluation has been used as either an expert-based or user-based UEM for collecting usability problems either in laboratory-based or in remote-based usability evaluations (Hartson, Castillo, Kelso, Kamler, & Neale, 1996; Thompson & Williges, 2000). In addition, Neale, Dunlap, Isenhour, and Carroll (2000) have developed a collaborative critical incident procedure that requires dialogue between the user and the expert evaluator in order to enrich the usability problem specification. Translating and classifying the critical incidents into usability problem domains is still somewhat problematic.

4.3.2. Severity Ratings

The concept of realness of a candidate usability problem was introduced as a way to distinguish trivial usability problems from important ones in UEM studies. While this binary test of problem impact is necessary, it is not sufficient. A usability problem judged to be real can still have either only minor impact on user satisfaction or it might have show-stopping impact on user task performance. To further discriminate among degrees of impact, practitioners have extended the binary concept of realness into a range of possibilities called severity levels. Severity thus becomes another measure of the quality of each usability problem found by a UEM, offering a guide for practitioners in deciding which usability problems are most important to fix. The working assumption is that high severity usability problems are more important to find and fix than low severity ones. Thus, a UEM that detects a higher percentage of the high severity problems will have more utility than a UEM that detects larger numbers of usability problems but ones that are mostly low severity (even though all problems found might be "real" by the definition used). There are numerous schemes for subjectively determining severity ratings for usability problems. Nielsen (1994) is a representative example. Rubin (1994) uses a criticality rating combining severity and probability of occurrence.

5. UEM Performance Measures – Applying Actual Criteria

Bastien and Scapin (1995) identified three measures for examining an evaluation method: validity, thoroughness, and reliability. Sears (1997) also points out these same measures, giving them somewhat different operational definitions. These basic three measures are:

- *Reliability:* Evaluators want consistent UEM results, independent of the individual performing the usability evaluation.
- *Thoroughness*: Evaluators want results to be complete; they want UEMs to find as many of the existing usability problems as possible.

- *Validity*: Evaluators want results to be "correct;" they want UEMs to find only problems that are real.

As a practical matter, we would add a metric we call *effectiveness*, which is a combination of thoroughness and validity. Additionally, on behalf of practitioners who must get real usefulness within tightly constrained budgets and schedules, we would also hasten to add *downstream utility* and *cost effectiveness*.

As Gray and Salzman (1998) point out, there is a need for multi-measure criteria, not just one-dimensional evaluations. When a researcher only focuses on one measure (e.g., thoroughness), it is unlikely that this one characteristic will reflect overall effectiveness of the UEM. For example, if an inspection method focuses on high reliability, it does not guarantee that the output of an inspection will produce quality problem reports that communicate problems and causes precisely and suggest solutions for down-stream redesign activities. In addition to thoroughness and validity, researchers may also be interested in reliability, cost effectiveness, downstream utility, and usability of UEMs. Any of these issues could form the criteria by which researchers judge effectiveness. Although it is nearly impossible to maximize all of the parameters simultaneously, practitioners must be aware that focusing on only one issue at the expense of others can lead to an actual criterion having significant criterion deficiency.

First, the ultimate criteria must be matched to the goals of evaluation. The main goal addressed by UEM evaluation is to determine which UEM is "best." Beyond that, we ask, "Best for what?" Ultimate criteria should be selected with this more specific question in mind. In effectiveness studies of UEMs, the objective should then be to find those measures comprising actual criteria to best relate them to the ultimate criteria. Thus, the measures are a way of quantifying the question of how well a UEM meets the actual criteria.

5.1. Thoroughness

Thoroughness is perhaps the most attractive measure for evaluating UEMs. Sears (1997) defines thoroughness as a measure indicating the proportion of real problems found using a UEM to the real problems existing in the target interaction design:

$$Thoroughness = \frac{\text{number of real problems found}}{\text{number of real problems that exist}} \quad (1)$$

For example, if a given UEM found only ten of the 20 real usability problems that were determined to be in a target system (by some criterion yet to be discussed), that UEM would be said to have yielded a thoroughness of 10/20 = 0.5. UEMs with low thoroughness waste developer resources by leaving important usability problems unattended after investment in the usability evaluation process.

Whatever method is used to determine realness, that method can also be considered a UEM, in this case a definitional UEM_A (A referring to "actual criteria") that, however arbitrarily, determines realness. The output of this so far undefined UEM is considered the "perfect" yardstick against which other UEMs are compared. When applied to the target interaction design, UEM_A produces a definitional usability problem set, A, defining those real problems that exist in the design.

If P is the set of usability problems detected by some UEM_P being evaluated, then the numerator for thoroughness of UEM_P is computed by an intersection as in this equation:

$$Thoroughness = \frac{|P \cap A|}{|A|} = \frac{|P'|}{|A|} \quad (2)$$

where |X| is the cardinality of set X and P' is the set of *real* usability problems found by UEM_P.

Weighting thoroughness with severity ratings provides a measure that would reveal a UEM's ability to find all problems at all severity levels. Such a measure can be defined by starting with the definition of thoroughness of equation (1):

$$Thoroughness = \frac{\text{number of real problems found}}{\text{number of real problems that exist}} \quad (3)$$

and substituting weighted counts instead of simple counts of problem instances:

$$Weighted\ Thoroughness = \frac{\sum s(rpf_i)}{\sum s(rpe_i)} \quad (4)$$

where s(u) is the severity of usability problem u, rpf_i is the ith real problem found by the UEM in the target system, and rpe_i is the ith real problem that exists in the target system. This kind of measure gives less credit to UEMs finding mostly low severity problems than ones finding mostly high severity problems.

However, for many practitioners who want UEMs to find high severity problems and not even be bothered by low severity problems, this kind of thoroughness measure does not go far enough in terms of cost effectiveness. For them, perhaps the breakdown of thoroughness at each level of severity is better:

$$Thoroughness(s) = \frac{\text{number of real problems found at severity level (s)}}{\text{number of real problems that exist at severity level (s)}} \quad (5)$$

Practitioners will be most interested in thoroughness for high levels of severity (high values of s) and can ignore thoroughness for low severity. Or a measure of the average severity of problems found by a given UEM, independent of thoroughness, might be more to the point for some practitioners:

$$s_{avg}(UEM_A) = \frac{\sum s(rpf_i)}{\text{number of real problems found by } UEM_A} \quad (6)$$

and this could be compared to the same measure for other UEMs or to the same measure for the problems existing in the target system:

$$s_{avg}\ (exist) = \frac{\sum s\ (rpe_i)}{\text{number of real problems that exist}} \tag{7}$$

The above definition would identify UEMs good at finding the most important problems, even ones that do not score the highest in overall thoroughness.

If researchers believe severity is important enough, they can include it in the ultimate and actual criteria as another way to enhance the criterion definition. By including severity, researchers introduce the problem of finding an effective actual criterion that captures "severity-ness," since no absolute way to determine the "real severity" of a given usability problem exists.

Researchers planning to use severity ratings as part of the criteria for comparing UEMs, however, should be cautious. Nielsen (1994), using Kendall's coefficient of concordance, found inter-rater reliability of severity ratings so low that individual ratings were shunned in favor of averages of ratings over groups of inspectors. Nielsen then used the Spearman-Brown formula for estimating the reliability of the combined judgments and found the group ratings more reliable.

5.2. Validity

In general terms, validity is a measure of how well a method does what it is intended to do. Sears defines validity as a measure indicating the proportion of problems found by a UEM that are real usability problems:

$$Validity = \frac{\text{number of real problems found}}{\text{number of issues identified as problems}} \tag{8}$$

Validity and thoroughness can be computed using the same data, the usability problem sets generated, and the realness criterion. For example, a UEM that found 20 usability problems in a target system, of which only 5 were determined (by criteria yet to be discussed) to be real, would have a validity rating in this case of 5/20 = 0.25. UEMs with low validity "find" large numbers of "problems" that are not relevant or real, obscuring those problems developers should attend and wasting developer evaluation, reporting, and analysis time and effort.

As above, computing validity in terms of sets, we get:

$$Validity = \frac{|\ P \cap A\ |}{|\ P\ |} = \frac{|\ P'|}{|\ P\ |} \tag{9}$$

In Section 4.1.2.4, we claimed that the union of usability problem sets produced by all UEMs being compared is flawed as a standard set of existing usability problems. Here we show why. The following steps describe how a union set is built from a set of individual usability problem sets:

Look at each candidate problem in each individual set, one at a time.

1. Determine if candidate problem is "real" (per criterion) and discard if not real.
2. If real, compare candidate problem to each of those in union set.

3. If candidate is represented in set (already in union set in some form), possibly use some of wording of candidate problem to refine or improve wording of the problem that represents the candidate in the union set; discard candidate.
4. If candidate is not represented in union set, add to set.

This technique works better if the number of methods being compared is relatively large, increasing confidence that almost all the real problems have been found by at least one of the methods. But one negative effect of this approach is to eliminate validity as a metric, an effect we feel is important enough to all but preclude the union of usability problem sets as a viable approach, as we explain next.

Suppose a UEM comparison study was conducted to compare UEM_P, UEM_Q, and UEM_R. Let P(X) be the usability problem set found in interaction design X by UEM_P, and so on for UEM_Q and UEM_R. No standard UEM, UEM_A, is available, so the union of the output sets of the UEMs being evaluated is used as a substitute for the output of a standard, UEM_A:

$$A(X) = P(X) \cup Q(X) \cup R(X) \tag{10}$$

Even though most of these studies do not say so explicitly, they are using this union as the basis of an actual criterion. The number of usability problems in this union is bounded by the sum of cardinalities of the participating usability problem sets:

$$|A(X)| = |P(X) \cup Q(X) \cup R(X)| \leq |P(X)| + |Q(X)| + |R(X)| \tag{11}$$

The more UEMs participating in the comparison the more real usability problems will be included in the union usability problem set. Unfortunately, more non-real problems are also likely to be included, decreasing validity. However, this approach to an actual criterion, by definition, prevents any possibility of detecting the reduced validity. For example,

$$\textit{Validity of } UEM_P = \frac{|P(X) \cap A(X)|}{|P(X)|} \tag{12}$$

Because A(X) is a union containing P(X), P(X) is a proper subset of A(X) and nothing is removed from P(X) when it is intersected with A(X). Thus,

$$\textit{Validity of } UEM_P = \frac{|P(X)|}{|P(X)|} \tag{13}$$

which is identically equal to 1.0.

In other words, this approach guarantees that the intersection of the UEM usability problem set and the standard usability problem set (the union) will always be the UEM usability problem set itself. This means that all usability problems detected by each method are always real and validity is 100% for all participating methods!

5.3. Effectiveness

Thoroughness and validity have rather nice analogies to the concepts of recall and precision in the field of information storage and retrieval, terms that refer to measures of retrieval performance of an information system from a target document collection (Salton & McGill, 1983). The document collection searched by an information system corresponds to the target interaction design being evaluated and the information system corresponds to the UEM. Precision and recall are based on a concept called relevance (reflecting a determination of relevance of a document to a query), analogous to the concept of realness in UEMs. Relevance is the criterion for computing precision and recall.

Recall corresponds to thoroughness and is a measure indicating the proportion of relevant documents (for example) found in a collection by an information system to the total relevant documents existing in the target document collection.

$$\text{Recall} = \frac{\text{number of relevant documents found}}{\text{number of relevant documents that exist}} \tag{14}$$

Precision is the proportion of the documents retrieved by an information system that are relevant:

$$\text{Precision} = \frac{\text{number of relevant documents found}}{\text{total number of documents retrieved}} \tag{15}$$

Just as neither precision nor recall alone is sufficient to determine information system retrieval effectiveness, neither thoroughness nor validity alone is sufficient for UEM effectiveness. For example, high thoroughness alone allows for inclusion of problems that are not real, and high validity alone allows real problems to be missed. Following the concept of a figure of merit combining precision and recall in information retrieval, we propose to capture the simultaneous effect of UEM thoroughness and validity in a "figure of merit" that we call "effectiveness," which we define to be the product of thoroughness and validity:

$$\text{Effectiveness} = \text{Thoroughness} \times \text{Validity} \tag{16}$$

Effectiveness will have the same range of values as thoroughness and validity, from 0 to 1. Where either thoroughness or validity is low, effectiveness will be low also. The effectiveness measure offers a compromise target for optimization. There can also be more interesting UEM design factors, beyond sensitivity, that might enhance one measure specifically without a negative effect on the other. We conjecture, in Figure 5 about the relationship among thoroughness, validity, and effectiveness.

M_1 A

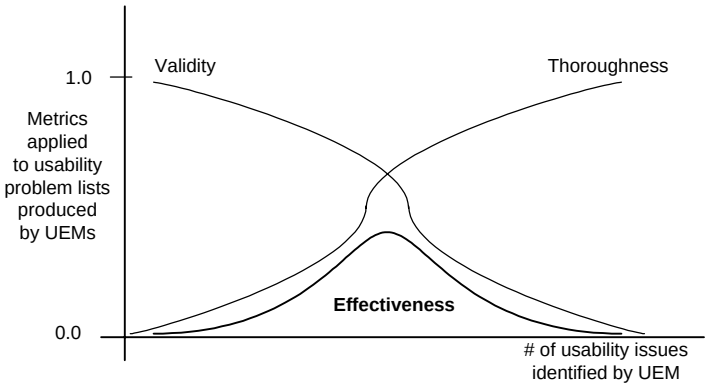

Figure 5. Conjecture About Relationship Among Thoroughness, Validity, and Effectiveness.

Several authors, such as Gray and Salzman (1998), have alluded to the concepts of hits, misses, false alarms, and correct rejections in the context of UEM outputs, concepts closely related to thoroughness, validity, and effectiveness. These concepts originated with hypothesis testing error types explained in most modern books on statistics or experimental research (for example, Keppel (1991) or Winer, Brown, Michels (1991)) and adapted for signal detection theory (Swets, 1964; Egan, 1975). Further adapting this terminology to usability problem detection, we can identify four cases, as shown in Figure 6, to describe the accuracy of a UEM with respect to the realness of the problems it detects, as determined by some actual criterion, A.

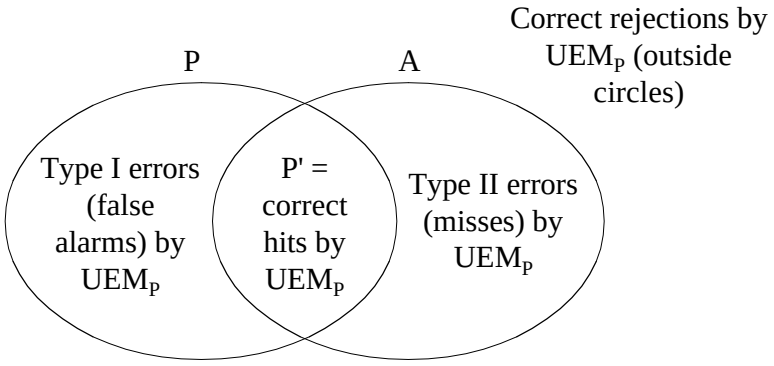

Figure 6. Venn Diagram of Comparison of a UEM_P Usability Problem Set Against Actual Criterion set, A.

High thoroughness is achieved in a UEM by realizing most of the correct hits and by avoiding most of the Type II errors (misses). Similarly, high validity of a UEM derives from avoiding most of the Type I errors (false alarms) and realizing most of the correct rejections. False alarms do not affect thoroughness, but do detract from validity. The intersection in the center and the area outside both ovals represents the areas of highest effectiveness, where the UEM is in agreement with the actual criterion. The same diagrammatic technique could be used to compare usability problem sets of two different UEMs, where the left hand oval is the usability problem set, P, of UEM_P and the right hand oval is the set Q of method UEM_Q.

In their discussion about how well the output of a UEM matches the "truth" about real problems existing in a given target system, Gray and Salzman (1998) seem to miss the importance of actual criteria. They say that, because we do not have access to "truth," a diagram like Figure 6 is

misleading. But that is exactly why we use actual criteria in experimental design. The point of having actual criteria is to isolate this question of "truth" to be an issue only between the ultimate criterion and the actual criterion and restrict it to the domain of actual criteria selection. Once a suitable actual criterion is established, it stands in stead of the ultimate criterion ("truth") and becomes the standard for determining realness of usability problems for the study. The study can be performed without the need to look beyond the actual criterion for truth. If a better approximation can later be found to the truth of the ultimate criterion, the actual criterion can be updated and studies can be repeated.

5.4. Reliability

Reliability of a UEM is a measure of the consistency of usability testing results across different users (developers). Usually it is desirable that UEM output be independent of who is using the UEM. Pearson's r is an index that describes the extent to which two sets of data are related. This has been used as a measure of reliability in the context of usability problem sets (Nielsen, 1994). We find agreement more useful than correlation for most situations requiring a reliability measure.

As a formal measure, reliability is an index of agreement between two or more sets of nominal identification, classification, rating, or ranking data. Cohen's (1960) kappa is one example of a reliability measure. Kappa is a measure of the proportion of agreement beyond what would be expected on the basis of chance. Kappa has an approximately normal distribution and can be used to test the null hypothesis of no agreement beyond the chance level. Cohen's original kappa ranged from –1 to +1, but negative values for kappa do not correspond to reality in our application, where kappa is, therefore, scaled between 0 and 1, with 0 corresponding to only chance agreement and 1 corresponding to perfect agreement. While the original concept of kappa is limited to assessing agreement between two subjects, an extension (Fleiss, 1971) permits measuring agreement among several subjects. The extension also produces a kappa value between 0 and 1 and allows testing for agreement by reference to the normal distribution.

There are other ways to compute a reliability measure. Sears (1997) measures reliability by using the ratio of the standard deviation of the number of problems found to the average number of problems found. Nielsen (1994) used Kendall's coefficient of concordance to assess agreement among evaluators making severity ratings.

Although it is usually desirable for UEMs results to be consistent or reliable across different individual users, the goals for developing a UEM and the ecological validity of a study for evaluating it, will depend on how the UEM is used. Because a UEM typically gives low thoroughness for an individual inspector (e.g., an approximate average of 30% (Virzi, 1990; Wright & Monk, 1991; Virzi, 1992; Nielsen, 1994)), UEMs are usually applied by a group of inspectors and the individual results merged (Nielsen, 1994). If that is how a UEM is used in practice, a realistic comparison study of such methods should be based on group results. Individual results might still be of interest in understanding and tuning the method, but method performance measures (and method cost) should be represented by its application within groups.

In most UEMs low individual reliability allows the merging process described above to pay off in substantially higher group thoroughness. That is, it is the high variability across individual inspectors that gives breadth to the union of results. Thus, just as for the thoroughness measure, for purposes of comparing UEMs reliability should be taken across groups rather than across individuals.

This is also a case that illustrates how UEM developers must approach their goals with care. It is reasonable for tool developers to aspire to improve individual inspector reliability by "standardizing" the inspection process, but this must be done carefully. It could have the undesired side effect of *reducing group thoroughness*. If the UEM (particularly, inspection method) designers achieve higher individual reliability by narrowing the view of inspectors to some "standard" guidelines or heuristics and a "standard" way to apply them, in effect pointing all the inspectors down the same path, it could result in lower group thoroughness by cutting off the broadening effect of individual variation. Thus, it is probably better for UEM developers to strive for higher thoroughness first and often reliability, at least group reliability, will improve as well in the process.

5.5. Downstream Utility

John and Marks (1997) stand almost alone in their consideration in a UEM study of downstream utility of UEM outputs, which depends on the quality of usability problem reporting. John and Marks make the case for persuasiveness of problem reports, a measure of how many problems led to implemented changes. They also suggest evaluating the downstream ability of UEM outputs to suggest effective redesign solutions through usability testing of the redesigned target system interface. This approach has the laudable objective of finding the UEMs that add value or utility in the change process, but inclusion of a more extensive process with high variability brings into question the feasibility of a controlled study of this effect. The iterative cycle of interaction design and redesign is anything but a well-specified and consistent process, depending greatly on team and individual skills, experience, and project constraints. Also, the quality of problem reports is not necessarily an attribute of just the UEM. Many UEMs are designed to detect usability problems but problem reporting is left to the developers/evaluators using the UEM. In such cases, problem report quality will vary greatly according to the skills of the individual reporter at communicating complete and unambiguous problem reports. Further, usability practitioners do not usually fix all problems found by UEMs. A process leading to fixing the wrong problems, even if with high quality fixes, might not be most cost effective. We have to conclude that adding more of the iterative development cycle does not add significant value to UEM comparison.

We do, however, believe in exploring downstream utility and usability problem report quality as important facets of the overall usability engineering process. We suggest, however, separating these effects from the issues of primary UEM performance (e.g., thoroughness and validity), treating UEMs as functions that produce usability problem lists and treating the connection of these problem lists back to redesign as a separate process. Researchers who have studied usability problem extraction, description, and classification (Lavery, Cockton, & Atkinson, 1997; Hartson, et al., 1999) have done just that: regarding these as separate function of usability engineering support tools for classification and reporting, used in conjunction with UEMs. This separation also derives from the fact that most methods for problem description and reporting are independent of the evaluation method and can be used with any UEM.

Perhaps an alternative way to evaluate post-usability-testing utility of UEM outputs is by asking real-world usability practitioners to rate their perceptions of usefulness of problem reports in meeting the interaction development cycle needs of analysis and redesign within their own real development environment. Although this approach might be less costly than that of John and Marks and might lend additional ecological authenticity, it would have to be validated by the kind of study they reported.

5.6. Cost Effectiveness

Just as the ultimate criterion for evaluating a military plane depended on its performance in real battles under real battlefield conditions, we have defined the ultimate criterion for UEM performance in terms of how well it detects or identifies real usability problems in real interaction designs. For many that is the end of the story, but the people who make buying decisions for military planes know there is more. To them, the real criterion for the airplane is to win battles, but to do so at the lowest cost in dollars, human lives, and collateral damage. We see the same story in usability practice. For practitioners, the goal is to find real usability problems, to do so with the maximum effectiveness, and to do it at the lowest cost possible. To capture this practical consideration, we include cost (e.g., cost to learn and cost to use a UEM) as a metric. Combining cost with our effectiveness metric also yields cost effectiveness, a measure of efficiency.

Good choices for actual criteria would then take efficiency into account. Of course, efficiency, in terms of cost and performance, must be defined quantifiably to be compared. As an example, one can combine our effectiveness measure in a quotient with cost to yield cost effectiveness of UEM usage. Cost can be measured, for example, as a function of method or tool use, including the fixed overhead of learning a method or tool combined with variable time and effort of applying it. Perhaps the biggest difficulty in getting a measure one can have confidence in is in estimating cost quantitatively and doing it accurately and consistently.

6. Review of UEM Studies

In the spirit of the Gray and Salzman, (1998) review of UEM comparison studies, our own research interests led to explore the relative benefits of various experimental techniques used in UEM comparison studies. We initially approached UEM comparative studies through the use of meta-analysis techniques, attempting to accumulate experimental and correlational results across independent studies. For this effort, we used several criteria to ensure that selected studies were within the focus of the meta-analysis. The criteria we used were:

- The study must involve software usability evaluation.
- A comparison must be made in this as a study of UEMs, using lab-based testing with users as a standard of comparison.
- Summary statistics must be reported in the study such that effect sizes can be calculated. Relevant statistics included: percentages of problems (of the total) detected by any one UEM (thoroughness) and validity scores.

As soon as we began the meta-analysis process, we realized that a significant majority of the comparison studies in the HCI literature on UEM effectiveness did not provide the descriptive statistics needed to perform a meta-analysis. This confirms the Gray and Salzman (1998) concern with statistical conclusion validity of five popular UEMs where formal statistical tests were often not included. In addition, many studies did not compare their results to a standard such as lab-based testing with users. UEM comparison studies varied significantly in terms of the criteria used to make comparisons. Criteria included measures that ranged from cost effectiveness to thoroughness, with only a few studies consistently using the same criterion for comparisons. This incompleteness and inconsistency present barriers to meta-analysis; perhaps symptoms of a field that is still very young. In the end, we had to relax our own criteria for identifying studies; resulting in a descriptive summary of key studies.

We were able to find 18 studies that we could identify as featuring a comparison of UEMs in terms of thoroughness or validity. The comparisons were usually among UEMs or among different usage conditions for a single UEM (e.g., applied to different software systems). The 18 studies we identified do not represent the entire population of UEM comparison studies. A full analysis of UEM comparison studies might yield another set of studies where the focus is on issues such as severity, experts vs. non-experts, teams vs. individuals, cost, guidelines vs. no guidelines, etc. However, the 18 studies we selected allowed us to make some conclusions based on the points put forth in the current paper. A summary comparison of these 18 studies is provided in Table 1, as adapted from Andre, Williges, and Hartson (1999).

Table 1. Summary of UEM Effectiveness Studies (codes explained at end of table).

Study	Methods (subjects)	Thoroughness	Validity	Notes
Bastien and Scapin (1995)	EC (10) NM (10)	EC > NM EC (M=89.9, SD = 26.2) NM (M=77.8, SD=20.7) p<.03		• NM = No method. Subjects just listed problems without a method guiding them. • Study provided mean, stddev, and p-values.
Bastien et al. (1996)	EC (6) ISO (5) NM (6)	EC > ISO/NM EC (M=86.2, SD=12.7) ISO (M=61.8, SD=15.8) NM (M=62.2, SD=13.8) p <.01		• Study provided mean, stddev, and p-values.
Beer et al. (1997)	CW (6) TA (6)	TA > CW p < .001		• TA > CW for major, minor, and cosmetic problems
Cuomo and Bowen (1992) Cuomo and Bowen (1994)	HE (2) CW (2) GR (1)	GR > HE > CW	CW > HE > GR CW (58%) HE (46%) GR (22%)	• Not Reported: mean, stddev, and p-values. • CW: Team approach.
Desurvire et al. (1992) Desurvire and Thomas (1993)	HE (3) CW (3) PAVE (3) UT (18)		HE > PAVE > CW HE (44%) PAVE (37%) CW (28%)	• Not Reported: stddev and p-values. • PAVE improved DV & NE performance.
Doubleday et al. (1997)	HE (5) UT (20)	HE > UT HE (86) UT (38)		• Not Reported: mean, stddev, and p-values. • 39% of UT problems not identified by HE. • 40% of HE problems not identified by UT.
Dutt and Johnson (1994)	HE (3) CW (3)	HE > CW		• Not Reported: percentage, mean, stddev, and p-values.
Jeffries et al. (1991)	HE (4) CW (3) GR (3) UT (6)	HE > CW/GR > UT HE (50%) CW (17%) GR (17%) UT (16%)		• Not Reported: stddev. • HE also found highest number of least severe problems. • CW & GR essentially used a team of 3 people, not individuals.
John and Marks (1997)	CA (1) CW (1) GOMS (1) HE (1) UAN (1) SPEC (1)	HE > SPEC > GOMS > CW > CA > CA > UAN HE (31%) SPEC (24%) GOMS (16%) CW (15%) CA (.08%) UAN (.06%)	CW > SPEC > GOMS > HE > CA/UAN CW (73%) SPEC (39%) GOMS (30%) HE (17%) CA/UAN (0%)	• Not Reported: stddev and p-values. • Validity here is the number of problems changed by developer.
John and Mashyna (1997)	CW (1) UT (4)		CW (5%)	• Not Reported: mean, stddev, and p-values. • Case study approach.
Karat et al. (1992)	IW (6) TW (6) UT (6)	UT > TW > IW p <.01		• Not Reported: percentage and stddev. • Walkthroughs essentially used Heuristics for evaluation. • Evaluated 2 different systems, but did not characterize the difference between the 2 systems.

Table 1 (continued)

Study	Methods (subjects)	Thoroughness	Validity	Notes
Nielsen and Molich (1990)	HE (various)	HE problems found: 20% to 51% (M)		• Not Reported: stddev and p-values. • Compared different systems using HE.
Nielsen (1990)	TA (36)		TA found 49% (M) of problems	• Not Reported: stddev and p-values.
Nielsen (1992)	HE (overall)	HE overall average across 6 systems was 35%		• Not Reported: stddev. • Nielsen collapsed 6 HE studies
Sears (1997)	HE (6) CW (7) HW (7)	HE > HW > CW (combining 4 or 5 evaluators) HW > HE > CW (combining 2 or 3 evaluators)	HW > CW > HE	• UT used to determine actual problems. • No mean or stddev reported for thoroughness or validity.
Virzi et al. (1993)	HE (6) TA (10) UT (10)	HE > TA > UT HE (81%) TA (69%) UT (46%)		• Not Reported: stddev and p-values.
Virzi (1990)	TA (20)		TA found 36% (M) of problems	• Not Reported: stddev and p-values.
Virzi (1992)	TA (12)	M=32%, SD=.14		• Reported overall detection rate for individuals.

Codes for Methods

HE = Heuristic Evaluation
CW = Cognitive Walkthrough
GR = Guidelines Review
UT = Usability Lab Test
HARDWARE = Heuristic Walkthrough
COGNITIVE AFFORDANCE = Claims Analysis
NM = No method
UAN = User Action Notation
SPEC = Reading the Specification
IW = Individual Walkthrough (essentially used Heuristics, not CW process)
TW = Team Walkthrough (essentially used Heuristics, not CW process)
TA = Thinking Aloud
PAVE = Programmed Amplification of Valuable Experts
EC = Ergonomic Criteria
GOMS = Goals, Operators, Methods, & Selection Rules

A majority of the UEM comparison studies (14) used the thoroughness measure for comparison. Examining the thoroughness studies in closer detail, we found 7 studies specifically using the heuristic evaluation technique as a comparison to other UEMs. The heuristic evaluation technique is reported as having a higher thoroughness rating in 6 out of these 7 studies (85.7%). Thus, a natural conclusion from the thoroughness criterion is that heuristic evaluation appears to find more problems than other UEMs when compared head-to-head, and such a conclusion is often reported in the literature with only a few exceptions. However, many of these studies use a somewhat loose definition of thoroughness, based only on a raw count of usability problems found. If these studies had used the tighter definition of thoroughness presented in this paper, where only real usability problems count toward the thoroughness measure, thoroughness results might have been somewhat, but probably not greatly, different.

Inclusion of the "realness" criterion in the validity measure would penalize the heuristic method for the relatively high rate of false alarms and low impact problems it was reported to identify. However, use of the validity measure in these studies was disappointing in general, offering mixed results in terms of identifying a particular UEM that might be more effective for finding

problems that impact real users. Many of the studies we reviewed did not explicitly describe how the validity measure was calculated, especially in terms related to lab-based testing with users.

Because very few studies provide the appropriate descriptive statistics, a robust meta-analysis was nearly impossible. Researchers counting "votes" may be able to realistically conclude that the heuristic evaluation method finds more problems than other UEMs. It would be more profitable to be able to conclude that a particular UEM has the highest effectiveness (per our definition in Section 5.3) and therefore finds just those problems that impact users in real work contexts. Such a conclusion can only be made when the criteria we use to measure effectiveness are relevant to the real work context, highlighting the need for the usability research community to consider carefully the criteria they use in UEM studies.

7. OTHER CONSIDERATIONS IN UEM STUDIES

7.1. UEM Comparison Experiments Should Focus on Qualitative Data Gathering Abilities

As we said in Section 1.4, formative evaluation is evaluation to improve a design and summative evaluation is evaluation to assess a design. UEMs are for gathering qualitative usability data for formative evaluation of interaction design usability. However, as we also mentioned in Section 1.4, some formative UEMs (e.g., laboratory-based usability testing) have a component with a summative flavor in that they also gather quantitative usability data. However, these quantitative data do not offer the hope of statistical significance required in summative evaluation.

The real work of UEMs is to support formative usability data gathering and iterative redesign in order to achieve an acceptable level of usability, and UEM studies should focus solely and explicitly on the ability to gather this qualitative data (e.g., usability problem sets). It is neither feasible nor useful to attempt to compare UEM quantitative data gathering abilities for two main reasons:

- Quantitative data gathered depends on usability goals, such as error avoidance in safety critical systems, walk-up-and-use performance and ease-of-learning in public installations, engagement in video games, or long-term expert performance in complex systems.
- Quantitative data gathering abilities are not an inherent part of any UEM. Quantitative data are collected by users of UEMs that employ user-based task performance (not a feature of many UEMs) and data collection is done more or less independently of the UEM. Most quantitative data represents such variables as tine-on-task, error counts, and user satisfaction scores, all measured in essentially the same way regardless of the UEM being used.

Unfortunately, there is confusion about this point. For example, Gray and Salzman (1998) cite, as a drawback of some UEM studies, that "It is not clear that what is being compared across UEMs is their ability to assess usability." But, in fact, because of the above two points, the ability to assess usability is not compared in UEM studies, because quantitative data gathering abilities are not comparable. None of the studies Gray and Salzman cite attempts to compare how well UEMs provide quantitative data. In fact, many of the UEMs studied (most inspection methods) do not even produce quantitative data.

Thus, bringing quantitative usability data into discussions of UEM comparisons can add confusion. An example of this potential confusion is seen in this statement (Gray & Salzman, 1998): "When an empirical UEM is used to compare the usability of two different interfaces on some measures(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)." First of all, per the above points, this statement is out of place in a discussion about experimental validity of UEM studies. In addition, "clear and unambiguous" comparison results can come only from a UEM being used for summative evaluation of the interfaces, but the context is about UEMs for formative usability evaluation.

Although qualitative usability data as collected by UEMs are central to UEM studies, readers of Gray and Salzman (1998) could be left with a feeling of ambiguity about their importance. In fact, Gray and Salzman say: "None of the studies we reviewed report systematic ways of relating payoff problems to intrinsic features; all apparently rely on some form of expert judgment." But that connection of user performance to causes in usability problems is precisely what one gets from qualitative data produced by a UEM. Usability problem lists and critical incidents or verbal protocol data provide relationships (attributions of cause) to intrinsic features for observed problems in payoff performance. Evaluators using a UEM may not always get the causes correctly, but each real usability problem in the output list is potentially the cause of *some* payoff problem.

7.2. Limitations of UEM Studies

Gray and Salzman (1998) propone the value of the experimental approach to provide strong tests of causal hypotheses, strong inferences about cause and effect. With proper experimental design for internal validity, one can be sure that a difference in the independent variable (the treatment) is, indeed, the cause of an observed difference in the dependent variable. The concern noted by Gray and Salzman regarding the strength of possible inference about causality is very difficult to resolve, however, in the case of UEM studies, where one is comparing one UEM against another that is potentially entirely different. The differences are far too many to tie up in a tidy representation by independent variables, forcing us to compare apples and oranges (Karat, 1998).

Monk (1998), in his capsule of the uses and limitations of experimental methods, points out that experimental studies are intended for deciding small questions among apples and apples, not grand cross-fruit questions such as deciding what UEM to use and why one is better. When the only independent variable that can be isolated in a "treatment" is the choice of UEM used, it means a black box view of the UEMs is being used, precluding the possibility of even asking the causality question, about *why* one UEM was better. Causality questions cannot go into more detail than the detail represented by the independent variables themselves.

The black box view of a UEM highlights the fact that the only thing the UEMs can be relied on to have in common is that they produce usability problem lists when applied to a target interaction design (the essential characteristic of a UEM pointed out in section 1.3). This fact heavily influences the possible choices for UEM performance measures. In essence, it means that an actual criterion to evaluate "how well" various UEMs produce these lists will have to be based on something that reveals the quality of those lists (e.g., realness), quantified within measures of that usability data.

Although the narrow black box view of UEMs does make comparisons of basic performance measures tractable, it does ignore other aspects and characteristics that could be important distinguishers between UEMs. It is appropriate, for example, to adjust the conditions for UEM application or to adjust the criteria for validity (realness) of usability problems detected. In other words, it is appropriate for UEMs with special talents to compete for attention in a correspondingly tailored arena, as long as (as Gray and Salzman have pointed out) the ground rules and conditions are the same for each UEM and are made clear in any report of the results. For example, if a given UEM

was good at focusing an evaluation instance to narrow the scope of the inspection to just meet evaluation needs at a certain stage of product development, that might represent a cost savings over another UEM that would force broader than necessary inspection at that same juncture in development. This would be a fair, albeit narrow comparison, of apples and apples, or at worst apples compared with oranges trying to act like apples – which is acceptable if practitioners do, in fact, need apples today.

7.3. Classification of Usability Problems for Structured Descriptions

Because classification aids description, some researchers are developing usability problem classification schemes to support more uniform problem description. As Gray and Salzman (1998) point out: One would think our discipline would already have a set of common categories for describing our most basic concepts in usability, but no standard categories yet exist. Gray and Salzman encountered three types of classification in the studies they reviewed: categories created in course of a study to account for data collected, lists of attributes from the literature, and one by Cuomo and Bowen (1994) based on theory. Lavery, Cockton, and Atkinson (1997) and Cockton and Lavery (1999) also found they needed a classification scheme in order to compare and match problems predicted by analytic methods to problems observed by empirical methods. Their framework for Structured Usability Problem EXtraction (SUPEX) is aimed at quality usability problem description and reporting through reliable problem extraction.

Like Cuomo and Bowen, we have followed the path blazed by Norman (1986) with his seven-stages-of-action model of interaction and have developed a detailed classification framework of usability attributes, called the User Action Framework (Hartson, et al., 1999), with which to classify usability problems by type. Often usability problems with underlying similarities can appear very different on the surface, and vice versa. The User Action Framework allows usability engineers to *normalize* usability problem descriptions based on identifying the underlying usability problem types within a structured knowledge base of usability concepts and issues. The User Action Framework provides a highly reliable (Andre, Belz, McCreary, & Hartson, 2000) means for a detailed classification of usability problems by a hierarchical structure of usability attributes, locating a usability problem instance very specifically within the usability/design space. The set of attributes, determined a node at a time along a classification path, represents a kind of "canonical encoding" of each usability problem in standard usability language.

7.4. Usability

We find it interesting that very little of the UEM evaluation literature, if any, mentions UEM usability as a UEM attribute or comparison measure. Perhaps we, as usability researchers and practitioners, should apply our own concepts to our own tool development. Surely UEM usability is a factor in determining its cost to use and, therefore, ought to be part of a criterion for selecting a UEM.

8. CONCLUSION

Although categories of UEMs are becoming somewhat well-defined in the HCI discipline, techniques for evaluating and comparing UEM effectiveness are not yet well-established. We believe it is possible to develop stable and consistent criteria for UEM effectiveness. Thoroughness, validity, and reliability appear to form the core of criterion measures researchers should continue to investigate. Thoroughness and validity measures must take into account the question of usability

problems realness and lab-based testing with users appears to be an effective way to provide a "standard" set of real usability problems. Although not an exact replication of real work contexts, user-based lab testing does provide a good indication of the types of problems that actually impact users. Another possibility for researchers is to push for examining problems that real uses do encounter in real world contexts using field studies and remote usability evaluation (Hartson & Castillo, 1998). The main difficulty with these methods, however, is that the lack of controls on tasks users perform can mean an inability to compare results among such UEMs.

We also believe that both usability researchers and usability practitioners will benefit from methods and tools designed to support UEMs by facilitating usability problem classification, analysis, reporting, and documentation, as well as usability problem data management (Hartson, et al., 1999). In the context of UEM evaluation we regard a reliable usability problem classification technique as essential for comparing usability problem descriptions, required at more than one point in UEM studies.

Finally, researchers should consider ways to reduce criterion deficiency and criterion contamination. We believe the easiest way to reduce criterion deficiency is through the use of several measures in the actual criterion, each focusing on a different characteristic of the UEM. In addition, it may be possible to examine how multiple measures can be combined into a composite measure that has a stronger relationship to the ultimate criteria.

At this point in the HCI field, it appears to be nearly impossible to do an appropriate meta-comparison of usability studies. We believe there are two reasons that contribute to the challenge of comparing UEMs. First, UEMs are extremely young (less than ten years) when compared to social science disciplines where baseline studies are frequently performed. Because of its youth, baseline comparative studies are still almost non-existent. Second, the methods for usability evaluation themselves are not stable. In fact, they continue to change because human-computer systems, their interaction components, and their evaluation needs change rapidly, requiring new kinds of UEMs and constant improvement and modifications to existing UEMs.

It was our objective in this paper to help alleviate the problems of variation, incompleteness, and inconsistency in UEM evaluate and comparison studies. We urge careful consideration of comparison criteria, both by researchers who perform UEM evaluation and comparison studies and by practitioners who use those studies to understand the relative merits of particular UEMs. We think of these suggestions and definitions, not as the final word, but more as a point of departure for more discussion and collaboration in bringing more science to bear on UEM development, evaluation, and comparison.

9. REFERENCES

Andre, T. S., Belz, S. M., McCreary, F. A., & Hartson, H. R. (2000). Testing a framework for reliable classification of usability problems. *Proceedings of Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society, to appear.

Andre, T. S., Williges, R. C., & Hartson, H. R. (1999). The Effectiveness of Usability Evaluation Methods: Determining the Appropriate Criteria. *Proceedings of Human Factors and Ergonomics Society 43rd Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society, 1090-1094.

Bastien, J. M. C., & Scapin, D. L. (1995). Evaluating a user interface with ergonomic criteria. *International Journal of Human-Computer Interaction, 7*(2), 105-121.

Bastien, J. M. C., Scapin, D. L., & Leulier, C. (1996). Looking for usability problems with the ergonomic criteria and with the ISO 9241-10 dialogue principles. *Proceedings of CHI Conference on Human Factors in Computing Systems*. New York: ACM Press, 77-78.

Beer, T., Anodenko, T., & Sears, A. (1997). A pair of techniques for effective interface evaluation: Cognitive walkthroughs and think-aloud evaluations. *Proceedings of Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*. Santa Monica, CA: Human Factors & Ergonomics Society, 380-384.

Bias, R. (1991). Walkthroughs: Efficient collaborative testing. *IEEE Software, 8*(5), 94-95.

Bradford, J. S. (1994). Evaluating high-level design: Synergistic use of inspection and usability methods for evaluating early software designs. In Nielsen & Mack (Eds.), *Usability Inspection Methods.* New York: John Wiley & Sons, 235-253.

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

Carroll, J. M., Singley, M. K., & Rosson, M. B. (1992). Integrating theory development with design evaluation. *Behaviour and Information Technology, 11*(5), 247-255.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *Proceedings of CHI Conference on Human Factors in Computing Systems*. New York: ACM, 213-218.

Cockton, G., & Lavery, D. (1999). A Framework for usability problem extraction. *Proceedings of INTERACT '99*. London: IOS Press, 344-352.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cuomo, D. L., & Bowen, C. D. (1992). Stages of User Activity Model as a Basis for User-Centered Interface Evaluation. *Proceedings of Annual Human Factors Society Conference*. Santa Monica: Human Factors Society, 1254-1258.

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

del Galdo, E. M., Williges, R. C., Williges, B. H., & Wixon, D. R. (1986). An Evaluation of Critical Incidents for Software Documentation Design. *Proceedings of Thirtieth Annual Human Factors Society Conference*. Anaheim, CA: Human Factors Society, 19-23.

del Galdo, E. M., Williges, R. C., Williges, B. H., & Wixon, D. R. (1987). A critical incident evaluation tool for software documentation. In L. S. Mark, J. S. Warm, & R. L. Huston (Eds.), *Ergonomics and Human Factors.* New York: Springer-Verlag, 253-258.

Desurvire, H. W., Kondziela, J. M., & Atwood, M. E. (1992). What is gained and lost when using evaluation methods other than empirical testing. In A. Monk, D. Diaper, & M. D. Harrison (Eds.), *People and computers VII.* Cambridge, UK: Cambridge University Press, 89-102.

Desurvire, H. W., & Thomas, J. C. (1993). Enhancing the performance of interface evaluators. *Proceedings of Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*. Seattle, WA: Human Factors and Ergonomics Society, 1132-1136.

Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. *Proceedings of Designing Interactive Systems (DIS '97) Conference*. New York: ACM Press, 101-110.

Dutt, A., Johnson, H., & Johnson, P. (1994). Evaluating Evaluation Methods. In G. Cockton, S. W. Draper, & G. R. S. Weir (Eds.), *People and Computers, Volume IX*. Cambridge, UK: Cambridge University Press, 109-121.

Egan, J. P. (1975). Signal detection theory and ROC analysis New York: Academic Press.

Ericsson, A., & Simon, H. (1984). Protocol analysis: verbal reports as data Boston: The MIT Press.

Flanagan, J. C. (1954). The Critical Incident Technique. *Psychological Bulletin, 51*(4), 327-358.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin, 76*, 378-382.

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.

Hartson, H. R., Andre, T. S., Williges, R. C., & van Rens, L. S. (1999). The User Action Framework: A Theory-Based Foundation for Inspection and Classification of Usability Problems. *Proceedings of the 8th International Conference on Human-Computer Interaction (HCI International '99): Ergonomics and User Interfaces, Vol. 1*. Mahway, NJ: Lawrence Erlbaum Associates, 1058-1062.

Hartson, H. R., & Castillo, J. C. (1998). Remote evaluation for post-deployment usability improvement. *Proceedings of AVI '98 (Advanced Visual Interfaces)*. L'Aquila, Italy:, 22-29.

Hartson, H. R., Castillo, J. C., Kelso, J., Kamler, J., & Neale, W. C. (1996). Remote Evaluation: The Network as an Extension of the Usability Laboratory. *Proceedings of CHI Conference on Human Factors in Computing Systems*. New York: ACM, 228-235.

Hix, D., & Hartson, H. R. (1993). Formative Evaluation: Ensuring Usability in User Interfaces. In L. Bass & P. Dewan (Eds.), *Trends in Software, Volume 1: User Interface Software*. New York: Wiley, 1-30.

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 CHI Conference on Human Factors in Computing Systems*. New York: ACM, 119-124.

John, B. E., & Marks, S. J. (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour and Information Technology, 16*, 188-202.

John, B. E., & Mashnya, 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), 1004-1022.

Kahn, M. J., & Prail, A. (1994). Formal usability inspections. In Nielsen & Mack (Eds.), *Usability Inspection Methods.* New York: John Wiley & Sons, 141-171.

Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. *Proceedings of CHI Conference on Human Factors in Computing Systems*. New York: ACM, 397-404.

Karat, J. (1998). The Fine Art of Comparing Apples and Oranges. *Human-Computer Interaction, 13*(3), 265-269.

Keppel. (1991). Design and Analysis: A Researcher's Handbook New York: Prentice-Hall.

Kies, J. K., Williges, R. C., & Rosson, M. B. (1998). Coordinating computer-supported cooperative work: A review of research issues and strategies. *Journal of the American Society for Information Science, 49*(9), 776-779.

Landauer, T. K. (1995). The Trouble with Computers: Usefulness, Usability, and Productivity Cambridge, MA: MIT Press.

Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology, 16*(4/5), 246-266.

Lewis, C., Polson, P., Wharton, C., & Rieman, J. (1990). Testing a Walkthrough Methodology for Theory-Based Design of Walk-up-and-Use Interfaces. *Proceedings of CHI Conference on Human Factors in Computing Systems*. New York: ACM, 235-242.

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

Lund, A. M. (1998). Damaged Merchandise? Comments on Shopping at Outlet Malls. *Human-Computer Interaction, 13*(3), 276-281.

Marchetti, R. (1994). Using usability inspections to find usability problems early in the lifecycle. *Proceedings of Pacific Northwest Software Quality Conference*. Palo Alto, CA: Hewlett-Packard, 1-19.

Meister, D., Andre, T. S., & Aretz, A. J. (1997). System Analysis. In T. S. Andre & A. W. Schopper (Eds.), *Human Factors Engineering in System Design.* Dayton, OH: Crew System Ergonomics Information Analysis Center, 21-55.

Monk, A. F. (1998). Experiments Are For Small Questions, Not Large Ones Like "What Usability Evaluation Method Should I Use?". *Human-Computer interaction, 13*(3), 296-303.

Neale, D. C., Dunlap, R., Isenhour, P., & Carroll, J. M. (2000). Collaborative critical incident development. *Proceedings of Human Factors and Ergonomics Society 43rd Annual Meeting,*. Santa Monica: Human Factors and Ergonomics Society.

Newman, W. M. (1998). On Simulation, Measurement, and Piecewise Usability Evaluation. *Human-Computer Interaction, 13*(3), 316-323.

Nielsen, J. (1989). Usability Engineering at a Discount. In G. Salvendy & M. J. Smith (Eds.), *Designing and Using Human-Computer Interfaces and Knowledge-Based Systems.* Amsterdam: Elsevier Science Publishers, 394-401.

Nielsen, J. (1990). Evaluating the thinking aloud technique for use by computer scientists. In H. R. Hartson & D. Hix (Eds.), *Advances in human-computer interaction.* Norwood, NJ: Ablex, 69-82.

Nielsen, J. (1992). Finding Usability Problems Through Heuristic Evaluation. *Proceedings of CHI Conference on Human Factors in Computing Systems*. New York: ACM, 373-380.

Nielsen, J. (1994). Heuristic Evaluation. In Nielsen & Mack (Eds.), *Usability Inspection Methods.* New York: John Wiley & Sons, 25-62.

Nielsen, J., & Mack, R. L. (Eds.). (1994). *Usability Inspection Methods*. New York: John Wiley & Sons, Inc.

Nielsen, J., & Molich, R. (1990). Heuristic Evaluation of User Interfaces. *Proceedings of CHI Conference on Human Factors in Computing Systems*. New York: ACM, 249-256.

Norman, D. A. (1986). Cognitive Engineering. In D. A. Norman & S. W. Draper (Eds.), *User Centered System Design.* Hillsdale, NJ: Erlbaum, 31-61.

Olson, G. M., & Moran, T. P. (1998). Commentary on "Damaged Merchandise?". *Human-Computer Interaction, 13*(3), 263-323.

Rubin, J. (1994). Handbook of Usability Testing New York: John Wiley & Sons, Inc.

Salton, G., & McGill, M. J. (1983). Introduction to Modern Information Retrieval New York: McGraw-Hill.

Scriven, M. (1967). The Methodology of Evaluation. In R. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of Curriculum Evaluation.* Chicago: Rand McNally, 39-83.

Sears, A. (1997). Heuristic Walkthroughs: Finding the Problems Without the Noise. *International Journal of Human-Computer Interaction, 9*(3), 213-234.

Smith, S. L., & Mosier, J. N. (1986). *Guidelines for Designing User Interface Software* (ESD-TR-86-278/MTR 10090). MITRE Corporation, Bedford, MA.

Swets, J. A. (1964). Signal detection and recognition by human observers New York: John Wiley.

Thompson, J. A., & Williges, R. C. (2000). Web-based collection of critical incidents during remote usability evaluation. *Proceedings of Human Factors and Ergonomics Society 43rd Annual Meeting,*. Santa Monica:.

Virzi, R., Sorce, J., & Herbert, L. B. (1993). A comparison of three usability evaluation methods: Heuristic, think-aloud, and performance testing. *Proceedings of Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting*. Seattle, WA: Human Factors and Ergonomics Society, 309-313.

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. *Proceedings of Human Factors and Ergonomics Society 34th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society, 291-294.

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

Wharton, C., Bradford, J., Jeffries, R., & Franzke, M. (1992). Applying Cognitive Walkthroughs to More Complex User Interfaces: Experiences, Issues, and Recommendations. *Proceedings of CHI Conference on Human Factors in Computing Systems*. New York: ACM, 381-388.

Winer, Brown, & Michels. (1991). Statistical Principles in Experimental Design New York: McGraw-Hill.

Wright, P., & Monk, A. (1991). A Cost-Effective Evaluation Method for Use by Designers. *Int. J. Man-Machine Studies*, *35*(6), 891-912.