Complementarity and Convergence of Heuristic Evaluation and Usability Test: A Case Study of UNIVERSAL Brokerage Platform

Lai-Chong Law
Computer Engineering and Networks Laboratory
Eidgenössische Technische Hochschule Zürich,
Gloriastr. 35, CH-8092 Zürich, Switzerland
+41 1 632 7837
law@tik.ee.ethz.ch

Ebba Thora Hvannberg
Systems Engineering Laboratory
University of Iceland
Hjardarhaga 2-6, 101, Reykjavik, Iceland
+354 525 4702
ebba@hi.is

ABSTRACT
The aim of this paper is twofold (i) comparing the effectiveness of two evaluation methods, namely heuristic evaluation and usability testing, as applied to an experimental version of the UNIVERSAL Brokerage Platform (UBP), and (ii) inferring implications from the empirical findings of the usability test. Eight claims derived from previous research works are reviewed with the data of the current study. While the complementarity and convergence of the results yielded by the two methods can be confirmed to a certain extent, no conclusive explication about their divergence can be obtained, especially the issue whether usability problems reported lead to failures in real use. One of the significant implications thus drawn is to conduct **meta-analysis** on a sufficient number of well-designed and professionally performed empirical works on usability evaluation methods.

Keywords
Heuristics evaluation, usability test, usability problem, brokerage platform

INTRODUCTION

Basic Concepts

Usability is defined as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use [11]. Similarly, a usability problem (UP) is defined as a flaw in the design of a system that makes the attainment of a particular goal with the use of the system ineffective and/or inefficient, and thus lowers the user's level of satisfaction with its usage. A variety of usability evaluation methods (UEMs) have been developed.

During the past decade, there have been a number of research studies comparing the effectiveness of different UEMs [1, 4, 6, 10, 12, 14, 15, 30]. Inconsistent empirical findings, however, are documented. The emergent trend is to delineate trade-offs - advantages and disadvantages - of individual UEMs [32] rather than to proclaim the relative effectiveness of different UEMs, given the controversial validity issues addressed critically by Gray and Salzman [9]. The results relevant to the present study are derived from those comparisons between heuristic evaluation (HE) and usability test (UT).

Briefly put, HE is a kind of analytic UEM conducted by a small group of evaluators, who examine a user interface, judge its compliance with a set of usability principles or heuristics, generate a list of usability problems (UPs), and, quite often, categorize the severity of UPs thus identified according to their estimated impact on user performance or acceptance. The recommended number of evaluators for a HE is between three and five, given that the informational gain with an additional evaluator drastically decreases after the fifth one and that the benefit-cost ratio is highest when three or four evaluators are employed [20]. In fact, among other forms of usability inspection method, HE is relatively more popular, primarily due to its ease of implementation and high efficiency.

UT is an empirical, time- and labour-intensive UEM involving a group of participants whose characteristics closely match those of real users of the product to be examined. During a test session, participants are usually required to think aloud while performing a set of test scenarios. The main benefit of the thinking aloud technique is a better understanding of the user's mental model and interaction with the product. One of the most important outcomes of UT is a list of UPs, which entail changes and improvement of the product. Based on the similar reasons identified for HE, the recommended number of participants in a UT is also five [24]. Nonetheless, this 'magic five' assumption can be challenged in view of other empirical findings [3, 17, 29].

In sum, the main distinguishing characteristic between HE and UT is that HE is easy to apply and quick to produce results, some of which may be irrelevant from the user's point of view, whereas UT is slow and laborious to apply but the results are accurate.

Product Description

The design of the product evaluated in the present study is based on an innovative concept – brokerage of e-Learning material. UNIVERSAL (http://www.ist-universal.org) is a European IST-Project with the primary goal to demonstrate an open exchange of learning resources (LR) between organizations whose members are registered users of the service. The organizations may be universities, business schools, training institutes or companies, and the individuals using the service are professors, researchers or training managers. The technology system that enables cataloguing, offers, enquiries, booking and delivery of a variety of LR, ranging from a short video to a complete live course, is known as UNIVERSAL Brokerage Platform (**UBP**). There are three main types of platform users: *providers* who offer LR, *consumers* who exploit LR for various purposes, and *administrators* who manage different types of accounts. An experimental prototype of the UBP was released in mid-October 2001, when the development work was somewhat at its middle stage. In fact, two structured functionality tests were performed on the earlier versions of the UBP. Thereby, a number of bugs with different levels of severity were fixed. The deployment of usability evaluation is a way to fine-tune the design.

AIMS AND SCOPE

With the primary goal to improve the usability of the UBP, empirical UT was our first choice, based on the assumption that it is the most effective usability evaluation technique. HE was additionally performed in order to uncover as many potential UPs as possible. The combined use of HE and UT is advocated, based on the arguments that they are complementary [21] and that HE is inherently limited [3]. The other rationale for such a synergy is that the results can allow us to investigate some of the claims about their relative effectiveness and respective trade-offs. Basically, we followed the standard procedures as described in the UEM literature but the number of evaluators involved in our HE was two instead of the recommended minimum of three. Nonetheless, this deviation, which was engendered by our limited resource, can enable us to address the issue about the optimal number of evaluators for HE with different arguments. In fact, Nielsen [22] found in his survey that the highest percent of the respondents employed two evaluators for HE, although they were instructed to use three to five. This shows that UEMs are normally adapted by practitioners to meet the specific conditions of the context to which they are applied. In addition, this 'localized approach' demonstrates so-called 'graceful degradation' in the sense that small deviations from the recommended practice only leads to slightly reduced benefits (see also our RESULTS below).

Note also that the type of software system we evaluated is relatively new in the European context and relatively culture-sensitive (e.g., one of the UPs identified by UT is multilinguality), the problem of generalization is thus apparent. Indeed, with the ever-increasing joint research and development projects across the European countries, the role of cultural differences in the perception of usability must seriously be addressed. In the present study, only two cultures, Swiss and Icelandic, were involved. We are somewhat convinced that more interesting findings would be obtained if more sites participated in the study.

PREVIOUS COMPARISON RESULTS

Among the related claims based on the previous HE-UT comparisons [3, 6, 12, 15, 21, 25, 30, 32], our data can validate and answer the following ones:

1. HE is more **cost-effective** than UT primarily in terms of shorter time period, from the conception of the test through its implementation to the release of the list of UPs, and the higher percentage of UPs, especially the minor ones.
2. HE and UT can **identify distinct sets** of UPs, and therefore they complement each other rather than lead to repetitive findings. Nevertheless, some studies emphasize that the results of HE and UT converge.
3. UT yields more **accurate** and **objective** results than HE. HE is likely to misidentify UPs (i.e. 'false alarms')
4. HE examines **intrinsic features** and attempts to make predictions concerning payoff performance. UT typically attempts to measure **payoff (or cost) performances** directly (e.g., speed, number of errors), which can be traced back to intrinsic features.
5. HE is more constrained than UT in terms of their relative pool of **evaluators**; for HE, usability specialists equipped with domain-expertise of a product are best possible candidates; for UT, general population with characteristics resembling real users of the product are appropriate candidates.
6. HE is more likely to **fail to identify positive features** of the product than is UT.
7. In general, usability inspection reports typically do not predict more than **30% to 50%** of end-user problem types.
8. When both HE and UT identify the same problem, does UT provide **deeper insight** into

the nature and origin of the problem than HE, or vice versa, or there is no gain of information when both instead of one are used?

The evaluation of the above claims with reference to our data will be discussed in a subsequent session.

METHOD AND PROCEDURE

Heuristic Evaluation

HE was performed on the experimental version of the UBP by the two evaluators (E1 and E2), who inspected the UBP independently. E1 is a computer scientist and usability specialist whereas E2 is a cognitive psychologist experienced in human-computer interaction. The heuristics used were taken from a set prepared by Molich and Nielsen [19], supplemented by three of Shneiderman's "eight golden rules of interface design" [26], which are distinct from the former set. The list of UP identified by E1 was sent to E2, who aggregated it with hers. Specifically, overlapped UPs between the two lists were discarded and the level of severity of some UPs was adjusted. Note that E1's and E2's familiarity with the UBP are described as low and high, respectively. In fact, E2 had participated in a functionality test on the pre-experimental version of the UBP three weeks before the implementation of HE. Note that the experimental version of the UBP was cleared of the bugs identified by the functionality test.

Usability Test

UT was performed in parallel on the same version of the UBP used in HE, and was administered by E1 and E2 with their respective team. Five participants were recruited in each of the two academic institutions where E1 and E2 reside (i.e., University of Iceland and Swiss Federal Institute of Technology [ETH] Zurich). The ten participants include university professors, project managers, research assistants, and administrators (details see Table 1). Given the remote collaboration, a usability test booklet containing all the testing materials, i.e., 20 task scenarios, see Table 2, pre/post-test questionnaires and a set of detailed usability test guidelines (e.g., defining characteristics of test participants, task distribution, measurements, etc.) were developed in order to standardize the procedures in both test sites. The ten test sessions were conducted separately during the period from 19th October 2001 to 9th November 2001. During a test session, each test participant was required to use the UBP to perform a set of twelve task scenarios (out of the total twenty) and to think aloud while carrying them out. An experimenter was present in the test room throughout the session to provide any assist solicited, to observe the test participant's behaviour, and to record some performance measures. Test sessions were videotaped and analyzed. The findings of individual sites were first compiled and interpreted locally and then combined and presented in an evaluation report [16], which was sent to the UNIVERSAL development team.

Table 1: Profile of the test participants of UT

Tester	Sex, Age	Job Title	ICT* Skills
P1	M, >50	University Professor	Very high
P2	F, >50	Administrator	Medium
P3	M, 30-39	University Professor	High
P4	F, 30-39	Research Assistant	Medium
P5	M, >50	Project Manager	Medium
P6	F, <30	Research Assistant	High
P7	F, 30-39	Project Manager	High
P8	F, 30-39	University Professor	Low
P9	M, 40-49	Technician	High
P10	M, 40-49	Administrator	High

Note: * ICT = Information and Communication Technologies

Table 2: List of 20 tasks in the usability test

Logging in
(T1) Logging in the UBP with username and password
User Management*
(T2) Registration of a new user
(T3) Modifying a user profile
(T4) Deleting a user from the UBP and logout
Higher Education Institution (**HEI**) Management*
(T5) Registration of a new HEI
(T6) Modifying a HEI profile
(T7) Deleting a HEI from the UBP and logout
Alliance Management*
(T8) Registration of a new alliance
(T9) Inviting HEIs to an alliance
(T10) Deleting an alliance from the UBP and logout
Learning Resource (**LR**) Provision*
(T11) Providing a new *packaged* learning resource
(T12) Modifying a *packaged* learning resource provided
(T13) Providing a new *live* learning resource
(T14) Modifying a *live* learning resource provided
(T15) Deleting the learning resource offer
Repository Access
(T16) Browsing the catalogue
(T17) Simple search for learning resources
(T18) Advanced search for learning resources
Booking and delivery
(T19) Booking & accessing a packaged learning resource
Assessment
(T20) Filling out the online UBP Questionnaire

Note: Those marked by asterisk (*) are tasks selectively performed by different participants.

MEASUREMENTS

The following subsections describe the measures taken in each of the tests.

Heuristic Evaluation

1. *Quantitative* measure: The number of UPs falling in each of the two severity levels: major vs. minor
2. *Qualitative* measure: Detailed descriptions of individual UPs and their locations (i.e. scope)

Usability Test

1. *Performance measures*, which were obtained primarily through observations. These measures concern counts of behaviours observed and consist of four aspects:
   a. Duration – Time to finish a task.
   b. Errors – (i) Number of wrong menu choices, (ii) number of wrong selections, and (iii) number of other errors.
   c. Seeking help – (i) Number of explicit requests for assist from the experimenter; (ii) Number of screens of online help looked at.
   d. Emotional expression – (i) Observation of frustration; (ii) Observation of confusion

   These measurements were collected with the use of a stopwatch and a paper log.
2. *Subjective measures*, which were obtained mainly through participants' self-reporting. These measures concern their perceptions, opinions, and judgments and consist of two aspects:
   a. Post-test Questionnaires: (i) Overall evaluation of the UBP (offline); (ii) UNIVERSAL Brokerage Platform Questionnaire (online)
   b. Thinking aloud protocols

RESULTS

Heuristic Evaluation

The two evaluators, E1 and E2, have independently prepared a list of UPs, which are partially overlapping. The divergences can be explicated with respect to their profile (see Table 3).

Table 3: Processing time of HE and profile of evaluators

	Processing Time	Familiarity With UBP	Usability Expertise	Domain* Expertise
E1	6 hours	low	high	medium
E2	3 hours	high	medium	medium

Note*: Domain expertise refers to knowledge and experience in brokerage of e-Learning material.

The short time expended by E2 could be attributed to the fact that she had participated in the earlier functionality tests of the UBP. Altogether, the two evaluators uncovered **43** UPs, with **25** and **18** of them being major and minor, respectively [16]. Breaking down these UPs into three groups, namely UPs uniquely identified by E1, UPs uniquely identified by E2, and UPs commonly identified by E1 and E2, reveals some interesting observations (see Table 4). E1 tended to identify more minor UPs, especially under the heuristic "consistency", whereas E2 tended to identify more major UPs, distributing in different heuristics. Indeed, such a distribution more or less resembles a typical contrast between HE and UT. It may be explained by the fact that E1's profile represents a typical HE evaluator whereas E2's profile represents an 'advanced user'. In addition, E2's previous exposure to the system may lower her sensitivity to minor problems and the functionality test in which she participated might serve as an exercise of informal feature inspection[1].

Table 4: Distribution of UP identified by two evaluators

	E1 (unique)	E2 (unique)	Common	Total
Major UP	3	13	9	25
Minor UP	9	3	6	18
Total	12	16	15	43

Usability Test

Performance measures

Each participant was required to perform a set of twelve tasks. The mean time-on-task for the whole set was 47.69 minutes (SD = 13.65), with the range from 31.46 (P4) to 77.62 minutes (P2). Note that for individual tasks we have set acceptable time ranges and number of errors, which have been calibrated with some participants in the functionality test. Altogether 120 tasks were performed, 97 were successfully completed, 8 were completed with assist (i.e., guidance/advice given by the experimenter) and 15 failed. We computed two usability metrics, namely effectiveness (i.e. un*assisted* mean completion rates and *unassisted* mean time-on-task per task) and efficiency (i.e. *unassisted* mean completion rate/ *unassisted* mean time-on-task). Table 5 lists the detailed results.

The mean effectiveness and mean efficiency of the twenty tasks over the ten participants are 75% and 48%, respectively. The three tasks with effectiveness less than 50% are Task 5 (creating a new HEI, 33%), Task 13 (providing a new live LR, 33%), and Task 14 (modifying the schedule of a live LR, 40%). Note that the mean completion times of Task 5 and Task 14 are almost double the upper bound of the acceptable range (cf. the given

[1] This inspection technique focuses on the function set of a product. Inspectors are usually given use cases with the expected result. Each function is analyzed for its availability, understandability, and other aspects of usability.

figures in square brackets), and the mean completion time of Task 13 exceeds the upper bound of the acceptable range by about 10%. These three tasks show the lowest efficiency with 2.95%, 2.00% and 4.38%, respectively. For effectiveness, we also looked at the total number of errors, number of assists, and frequency of expressing frustration per task (see Table 6).

Table 5: Effectiveness and efficiency per task

Task	Effectiveness		Efficiency
	Completion rate*	mean time #	
1	90% (100%)	1.29 [0.25-0.5]	69.53%
2	100%	4.07 [10 – 15]	24.57%
3	100%	1.12 [2 - 3]	89.29%
4	100%	0.78 [2 - 3]	128.21%
5	33% (100%)	11.17 [3 –5]	2.95%
6	67%	1.61 [2 – 3]	41.61%
7	67% (100%)	3.25 [2 –3]	20.62%
8	50%	2.82 [3 –5]	17.73%
9	50%	2.58 [3 – 5]	19.38%
10	100%	0.55 [2 - 3]	181.82%
11	60% (80%)	8.96 [10 -15]	6.7%
12	75%	2.59 [2 –3]	28.96%
13	33% (50%)	16.47 [10-15]	2.00%
14	40%	9.13 [3 - 5]	4.38%
15	70%	0.50 [2 –3]	141.21%
16	100%	3.49 [3 - 5]	28.65%
17	90% (100%)	1.16 [3 –5]	77.59%
18	100%	2.55 [5 –10]	39.22%
19	90% (100%)	3.75 [10 -15]	23.99%
20	90%	6.33 [10 -15]	14.23%

Note: * Assisted completion rate, which are different from the corresponding unassisted ones, are listed in parentheses.
The figures given in square brackets are the acceptable time ranges.

The mean error rate per participant over the twelve tasks performed is 11.8 (SD=6.56). In other words, on average they committed one error per task. The three tasks with the highest number of errors are Task 13, Task 16 (browsing the catalogue), and Task 11 (providing a new packaged LR). The correlation between the variables *Assist* and *Frustration* is moderately positive ($\rho = 0.485$). It implies that the higher the degree of frustration, the higher the tendency to seek assist is. However, this relationship is not particularly strong. Task 13 caused the highest frequency of frustration. Interestingly, for Task 1 (logging in), the frequency of frustration is relatively high, though the participants did not seek any assist. An interesting observation is that the test participants in one site tended to express more frustration than those in the other site, though the number of errors and assists were comparable between the two sites. The discrepancy could be attributed to the cultural difference in externalising emotion or it is an artefact generated by the experimenters' varied interpretations of facial and/or verbal expressions.

Table 6: Sum of errors, assists, and frustration per task

Task #	Errors*	Assists	Frustration
1 (10)	6 [1]	0	9
2 (5)	5 [2]	0	2
3 (5)	3 [2]	0	0
4 (5)	2 [2]	0	2
5 (3)	7 [2]	4	9
6 (3)	2 [1]	0	3
7 (3)	8 [1]	1	3
8 (2)	3 [2]	0	3
9 (2)	3 [2]	0	2
10 (2)	0 [1]	0	0
11 (5)	12 [3]	7	9
12 (4)	5 [1]	1	4
13 (6)	22 [3]	2	12
14 (5)	9 [1]	0	9
15 (10)	5 [1]	1	6
16 (10)	13 [1]	0	8
17 (10)	2 [1]	1	1
18 (10)	3 [2]	0	4
19 (10)	6 [3]	1	2
20 (10)	2 [2]	0	3

Note: # The figures in parentheses are the total number of test participants involved in the task
* The figures in square brackets are the acceptable number of errors per test participant per task.

The most problematic task is "Providing a live learning resource" (Task 13) (see Figure 1). Six participants performed it and all of them expressed considerable frustration and four sought assist from the interface help-text and the experimenter. Three of them failed to complete the task because they could not understand the terminologies and the structure of the online form. The range of the time is large, varying from 4.30 to 20.93 minutes. It can be accounted by the fact that one user gave up and the other employed 'trial-and-error' method. The range of the number of error is also large, varying from 1 (acceptable) to 8 (highly unacceptable). Indeed, similar observations were obtained for Task 11 and Task 14.

Subjective measures

Based on the participants' thinking-aloud protocols and experimenters' observations, **39** usability problems (UPs) were identified by ten participants of the two test sites. Of the 39 problems, there were **31** major and **8** minor, respectively [16]. The data consistently show that the three tasks under the functionality "Learning Resource Provision" have relatively high proportion of UPs: Task 11 (6 major, 1 minor), Task 13 (2 major, 1 minor), and Task 14 (3 major, 2 minor). In a subsequent session, we will compare the UPs uncovered in HE and UT.

Apart from effectiveness and efficiency, the third commonly employed usability metric is satisfaction. Here we gauged it with two questionnaires. In Task 20, the participants were required to click a link on the main page, navigating to the online UBP questionnaire that consists of 15 items of rating questions (5-point Likert scale). Eight participants partially or completely filled it out. Of particular interest is the item 13 that is used to evaluate the overall satisfaction using the UBP. The mean rate was 2.88 (n=8, SD=1.13), a bit above average. The participants were also asked to fill out a post-test questionnaire that consists of three open-end questions and two rating questions on overall ease of use and use-friendliness of the UBP with a 5-point Likert scale. The level of difficulty was perceived as moderate, with a mean of 2.3 (n=10, SD=0.823). The level of use-friendliness was perceived as satisfactory with the mean of 3.1 (n=10, SD=1.37). Considering that the UBP was still a prototype at the time of testing, such results were regarded as encouraging.

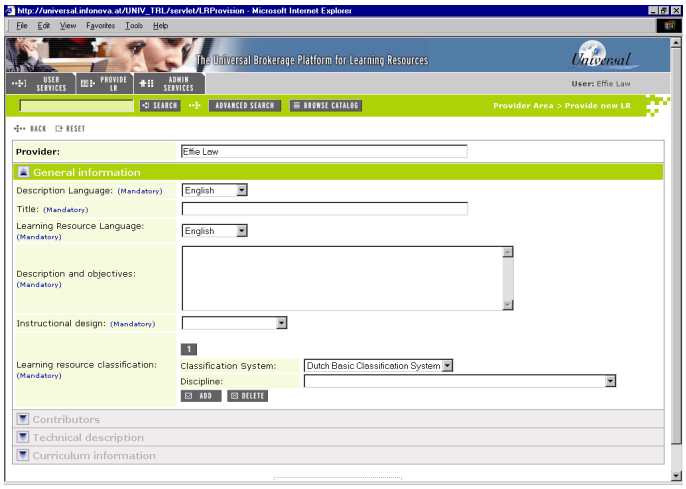

Figure 1: Screen dump of "Learning Resource Provision" function of the UNIVERSAL Brokerage Platform (Oct. 2001)

Comparisons of Usability Problems Identified

Here we compare the results of UT and HE with regard to the following questions: (i) How many *unique* major and *unique* minor UPs were identified by individual methods? (ii) How many *common* UPs were identified by both methods? (iii) What are the percentages of UPs identified by individual methods and individual evaluators? Note that we take a rough estimation that the total number of UPs equals to the sum of the unique UPs identified by HE, the unique UPs identified by UT and their common UPs. However, in reality, it is impossible to obtain a complete set of problems with an application [2, 13]. Nonetheless, some intriguing findings were obtained from the comparisons (Table 7 and Table 8).

Table 7: Unique and common UPs found by HE and UT

Problem Type	**Heuristic Evaluation (HE)**	**Usability Test (UT)**
Major = **40**	Common = 16 Unique = 9 Subtotal = 25	Common = 16 Unique = 15 Subtotal = 31
Minor = **23**	Common = 3 Unique = 15 Subtotal = 18	Common = 3 Unique = 5 Subtotal = 8

Table 8: Percentages of UPs found by four means

Total	**HE only**	**UT only**	**E1 only**	**E2 only**
63	43	39	27	31
100%	68.3%	61.9%	42.9%	49.2%

The percentages of UPs identified by HE and UT are 68.3% and 61.9%, respectively. On the face value, it seems that HE is more effective than UT. However, the fact that 15 minor UPs identified by HE are not 'verified' by UT may render this assumption dubious. It is not possible to tell whether these minor UPs are 'false alarm' or the sample of UT participants was unable to locate them [20]. Furthermore, E1 alone could identify 42.9% of the total number of UPs found by both methods. The additional gain by including E2's evaluation is 25.4% (i.e., 16 unique UPs, see Table 4). Both figures are higher than the average norms (see Figure 2).

Nielsen and Landauer [24] used the binominal probability formula $(1-(1-\lambda)^n)$ to calculate the number of evaluators or test participants required for HE or UT, where λ is the *proportion* of UP discovered when using a single evaluator or test participant, and n is the number of evaluators or test participants used. The typical values of λ for a single evaluator and a test participant are 0.34 and 0.31, respectively. Note that the figures computed are based on the proportion of the UPs identified by one evaluator (or one test participant) over the total of UPs found by all evaluators (or all test participants) in a particular study, designated by N. However, the number of UPs found by all evaluators (or test participants) is *not* the actual number of UPs in a system because of possible overlooking or misidentifying UPs.

Apparently, the total number of UPs derived from two UEMs, like what we have calculated here, is more accurate and usually larger than that based on one UEM. With a larger N, the value λ will become smaller than the typical norms. In other words, more evaluators or test participants than suggested by Nielsen [20] are required in order to yield a certain proportion of UPs. Interestingly, our data present some contradictory results. The pattern of the current UT, with ten test participants producing only 61.9% (cf. Nielsen's prediction: more than 95%), follows the proposed trend based on a smaller λ, whereas the pattern of the current HE goes the other way.

Nonetheless, it can be attributed to the two evaluators' special experiences with the system (cf. "*Wildcard effect*" [9], p.210]).

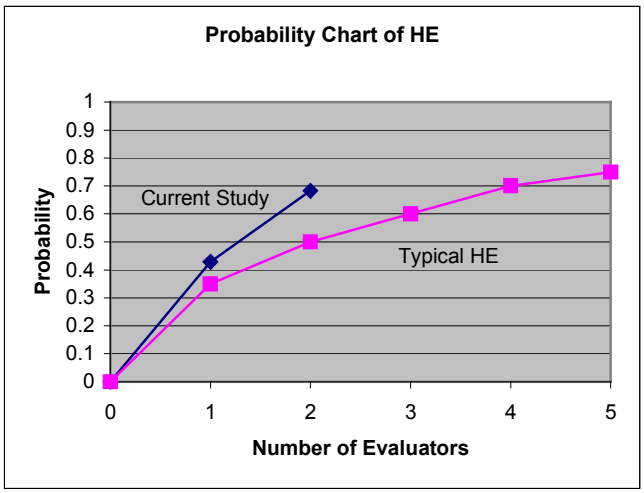

Figure 2: Probability charts of identifying UP in the current case study and typical HEs.

GENERAL IMPLICATIONS OF THE USABILITY TEST

The results of our UT not only enable the improvement of the system to be tested but also allow us to take closer look into the design of UT and conjoint use of UEMs. Here below we delineate some of the significant implications:

1. Usability problems identified by UT may not be related to the intrinsic features of the system, but to the flaws in the design of the task scenarios. For instance, the frequent shift of roles and the unduly repeated login/logout might have aggravated the confusion and frustration. Though the task scenarios have been pilot tested with a single user, the above problems were not identified. The involvement of the development team in the design of task scenarios may improve their quality.
2. The main problems identified in the prototype are typical usability problems, including lack of clear and timely feedback, navigation problem, lack of comprehensive help text, and excessive use of technical jargons and acronyms. Given that the designers and developers of the project are supposedly well informed about standard usability requirements for website design, the occurrence of so-called classic UPs brings forth the research concern how general usability requirements can effectively be translated and adapted to a specific practical context [31]. We advocate the recommended strategy of acquiring user requirements by involving them in design as early and frequently as possible. Nonetheless, the UNIVERSAL project is now working towards this direction by opening the UBP to potential end-users for public trials.
3. For the experiment's sake, we conducted HE and UT more or less in parallel. However, for maximizing their benefits, HE should be performed first and the usability problems thus identified should be fixed for the version to be examined by UT. Thereby, fewer participants are required to complete UT with a shorter period, i.e., the efficiency can be enhanced. Besides, the predictive power of HE can be evaluated if two small-scale UTs (instead of one large-scale) are to be performed before and after the fixing of the UPs. If there are improvements in performance measures, we may infer that HE is effective.

EVALUATING PREVIOUS COMPARISON RESULTS

In the foregoing literature review, we enumerate eight comparison results derived from the previous studies. Here we attempt to evaluate them with our data:

1. *Relative higher cost-effectiveness of HE*

   Nine hours and about 200 hours (each test session lasted on average 48 minutes) were spent in the design and conduction of HE and UT, respectively. HE found a higher percentage of UPs, exceeding UT by 6.4% or four UPs. UT located 6 more major problems than HE whereas HE located 10 more minor problems than UT. Moreover, the two evaluators of HE and the ten participants of UT found 68.3% and 61.9% of the total number of UPs, respectively. These findings are somewhat consistent with those of the previous studies.

2. *Convergence of results*

   Sixteen out of 40 major UPs (= 40%) and three out of 23 minor UPs (= 13%) were commonly found by both HE and UT. The convergence rates for both types of UPs are low. As it is relatively easier to identify major UPs than minor ones, it is not surprising that the test participants of UT found much less minor UPs than the two evaluators.

3. *Accuracy and objectivity of UT results and misidentification of problems in HE*

   Presumably, the UPs identified by test participants are accurate, because they represent or even will become real users of the system. How consistently should different test participants identify a UP so that it can be confirmed as a genuine UP? The idiosyncrasy of individual test participants, for instance, their technological knowledge and even personal aesthetic preference, affects whether a UP is named. In fact, some of the UPs reported in our UT were found by one (out of ten) test participant. Nonetheless, the problem of handling 'outliners' has been discussed in the related literature [10], but there

is no consistent view how it can be solved. Similarly, we cannot draw any definite conclusion about the issue of problem misidentification in the case of HE. The question is: Are the unique UPs identified in HE all false alarms, simply because they were not recognized by the selected (limited) sample of test participants? Concerning the objectivity issue, while there is no doubt that the quantitative measures like completion time and number of errors or assists are objective, we point out that such data will be only useful when they are compared with references, like the allowable time and error ranges we provided, or with a related set of measurements on a similar product.

4. *Linking intrinsic feature to payoff performance*

   Based on the list of descriptions of UPs identified by HE [16], we tend to agree with the claim that this analytical UEM is not apt for making "forward inference from intrinsic feature to payoff" ([9], p.216). Nonetheless, attempts have been made by E1 to find the origin of the usability problems by associating them with activities of a user interface development life cycle, from the initial phase of task analysis to the final phase of evaluation. This endeavour was demonstrated to be meaningful.

   As reported earlier, the tasks pertaining to Learning Resource Provision (Figure 1) were perceived to be most complicated. The development team of the UBP, considering the usability findings, has recently engaged in revising the learning resource taxonomy to render it more precise and concise, i.e., an exercise of conceptual re-modelling. The concomitant revision of navigational design is also in sight.

5. *Pool of evaluators vs. population of testers*

   Clearly, for HE, the limited availability of evaluators with relevant experience and knowledge is always a constraint, which is aggravated if the resource allotted to usability evaluation is restricted. In view of the lack of extra budget to employ 'external' reviewers, E1 and E2 assumed the dual roles – HE evaluator and UT administrator – for which they were well qualified. Though the time sequence of the two tasks with HE preceding UT could minimize the impact, such a practice of overlapping roles is not recommendable, considering the issue of validity [9]. As the UBP is an application with relatively broad target groups, there was no difficulty in recruiting the test participants, who took part in UT on a voluntary basis without payment.

6. *Positive findings*

   In HE no positive findings were reported. It is not surprising because the goal of such an exercise was problem finding. On the other hand, in UT the test participants were explicitly required to identify both positive and negative features of the system. Some positive comments were thus yielded.

7. *Predictive power of UEMs*

   There are two levels of this issue: How predictive is HE of usability problems found by sample test participants of UT? How predictive is laboratory-based UT of usability problems confronted by real end-user in actual working places? The former, which is a matter of empiricism (objectivity) vs. judgment (subjectivity), can be answered in terms of degree of overlapping of UPs identified by HE and UT. With the rate of 48.7% (i.e. 19 of 39 UPs found by UT could be predicted by HE), we conclude that the predictive power of HE is moderate. Nonetheless, the two questions are interrelated. If the predictive link addressed in the second question can be established, then it is more meaningful to investigate the link addressed in the first question.

   The second question actually tackles the tricky issue of external generalization. The problem hinges crucially on the degree of difference between the exact settings and persons used in the experiment and the wider range of settings and persons to which the experimental results are to be generalized. The question is how the change of setting will influence the behaviour of test participants: Are test participants more restrained to express their frustration or discontentment with a product when they are aware of being observed? Does their anxiety engendered by being put in a laboratory setting dampen or heighten their sensitivity to potential usability problems of the system, etc?

   Along the line of traditional experimental approach, the laboratory results are assumed to be representative of users' experience in the context of their work. However, the recent research in HCI, which is intimately related to the works in psychology, sociology, and anthropology [7, 28], has seriously challenged this assumption. One of the promising approaches advocated in this line of research is "participatory observation". Put briefly, end-users are observed unobtrusively *in situ* how they interact with a system in their working place, what kinds of problems they confront and how they resolve them. Further discussion on this intriguing topic, however, is beyond the scope of this paper.

8. *Accumulative insights into problems*

   With a close examination of the common UPs identified by both HE and UT, it is interesting to find that for the same UP the HE evaluators tended to describe it at a more general level whereas the UT test participants at a more detailed level. It may be attributed to the fact that the evaluators needed to go through all the functionalities whilst individual test participants focussed on a subset of them.

Presumably, the more exact the description of a problem is, the easier it is for the development team to understand the nature of the problem. On the other hand, neither groups tended to propose any solution for the UPs.

CONCLUSION

In the current study, we adopted the approach of multiple converging measures, which is highly advocated by a number of researchers in the domain of usability engineering. Not only have we drawn horizontal comparisons between an empirical usability testing (UT) and an analytic heuristic evaluation (HE), but we also drawn vertical comparisons among different behavioural measures within the usability testing. While the so-called general problem-type-based convergence across the two methods is relatively low, the specific task-based convergence as exemplified by the most problematic user interface of 'Learning Resource Provision' (LRP) is rather high. More specifically, about 44% of the usability problems identified in HE (19 out of 43) and, interestingly enough, the same percent in UT (17 out of 39) were found in the LRP pages. Task 13 of UT "Providing a Live Learning Resource" scored the lowest effectiveness, lowest efficiency, highest instances of errors, and highest frequency of frustration. The performance measures of the other related tasks (Task 11 and Task 14) were also rather poor. Definitely, convergent findings as such can strongly convince designers and developers to undertake corrective actions. Indeed, the UBP development team has already engaged in conceptual modelling of learning resource taxonomy, which eventually leads to re-design of LRP pages.

While complementarity and (partial) convergence of the results yielded by HE and UT can be verified by our data to a certain extent, our data do not enable us to provide any conclusive explanation about their divergence. If HE identifies a problem not found by UT, does it imply that evaluators of HE make a false alarm or test participants of UT overlook the problem? A chain of related questions can be raised. Among others, we highlight the following open questions: (i) How consistent a problem should be named by a sample of end-users so that it can be regarded as a genuine usability issue, 50% or any arbitrary proportion? (ii) How predictive is a laboratory-based UT of real usability problems arisen in an actual working place? (iii) Are the so-called standards valid, including the list of heuristics for evaluation [5, 8, 19, 21], the 'magic five' as the optimal number of evaluators and test participants, etc? While there are a handful of studies that attempted to validate the approach for estimating the optimal number of test participants or evaluators [2, 27, 29], systematic studies to compare different heuristics are scarce. It may be interesting to investigate the differences in the number and quality of usability problems identified when the same set of evaluators use different heuristics to assess the same or highly similar products.

Indeed, the research works in the field of HCI are broad in scope and diversified in topics. Whilst new research issues keep on emerging, some old ones remain unresolved. Specifically, in the area of usability engineering, the lack of a shared research context is a compelling concern, as exemplified by the striking observation that individual studies rarely replicate results from earlier studies [18, 24]. The current study, which nonetheless has its inherent limitation, has provided more empirical data to consider the validity of the claims about heuristic evaluation and usability testing. With the ever-increasing use of information and communication technologies, the role of usability engineering will become more critical. We foresee that many studies on usability evaluation methods will be performed in the coming years. We recommend that when a sufficient number of systematic and well-designed and professionally performed empirical works on usability evaluation methods are available, ***meta-analysis*** can be conducted on them to infer a clear, holistic, and more conclusive picture about this exciting field.

REFERENCES

1. Bailey, R.W., Allan, R.W., & Raiello, P. (1992). Usability testing vs. heuristic evaluation: A head-to-head comparison. *Proceedings of the Human Factors Society 36th Annual Meeting,* pp. 409-413.
2. Bailey, R.W. (2000). *Improving usability in America's voting systems. Calculating the number of test subjects required to find usability problems*. Available from: http://www.humanfactors.com/library/nov00.asp
3. Bailey, R.W. (2001). *Heuristic evaluations vs. usability testing.* Available from: http://www.humanfactors.com/library/jan012.htm
4. *Behaviour & Information Technology* (1997). Special issue on usability evaluation methods.
5. Cuomo, D.L., & Bowen, C.B. (1994). Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting with Computers, 6*, 86-108.
6. Desurvire, H.W. (1994). Faster, cheaper! Are usability inspection methods as effective as empirical testing? In J. Nielsen & R. Mack (eds.), *Usability inspection methods*, 173-201. New York: Wiley.
7. Dourish, P., & Button, G. (1998). On "Technomethodology": Foundational relationships between ethnomethodology and system design. *Human-Computer Interaction, 13*, 395-432.
8. Gerhardt-Powals, J. (1996). Cognitive engineering principles for enhancing human-computer performance. *International Journal of Human-Computer Interaction, 8*(2), 189-211.

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

10. *Human-Computer Interaction* (1998). Special issue on experimental comparisons of usability evaluation methods.

11. *ISO 9241: Guidance on usability standards*. Available from http://www.iso.ch/iso/en/CatalogueListPage.CatalogueList?ICS1=13&ICS2=180

12. 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 ACM CHI'91*, pp. 119-124. New York: ACM Press.

13. Jeffries, R., & Miller, J.R. (1998). Ivory towers in the trenches: Different perspectives on usability evaluations. *Human-Computer Interaction, 13*, 270-276.

14. Jorgensen, A. H. (1999). *Towards an epistemology of usability evaluation methods*. Available from: http://cyberg.curtin.edu.au/members/papers/43.shtml

15. Karat, C., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. *Proceedings of ACM CHI'92*, pp. 397-404.

16. Law, L.-C. (2002) (Ed.). *Trials Evaluation Report Y1, Deliverable D7A*, the UNIVERSAL project.

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

18. Mackay, W.E. (1998). Triangulation within and across HCI disciplines. *Human-Computer Interaction, 13*, 310-315.

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

20. Nielsen, J. (1994). Heuristic evaluation. In Nielsen, J., & Mack, R.L., (Eds.), *Usability inspection methods*, pp. 25-64. New York: John Wiley & Sons.

21. Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. CHI'94 Conference Proceedings.

22. Nielsen, J. (1995). *Technology transfer of heuristic evaluation and usability inspection*. Presented as a keynote speech as the IFIP INTERACT'95, Norway, June 27.

23.Nielsen, J. (2000). *Why you only need to test with 5 users*. Jackob Nielsen's Alertbox. Available from: http://www.useit.com/alertbox/20000319.html

24. Nielsen, J., & Landauer, T.K. (1993). A mathematical model of the finding of usability problems. *Proceedings of ACM/IFIP INTERCHI'93 Conference,* Amsterdam, the Netherlands, April 24-29, 206-213.

25.Oviatt, S.L. (1998). What's science got to do with it? Designing HCI studies that ask big questions and get results that matter. *Human-Computer Interaction, 13*, 303-307.

26. Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (3rd Ed.). Reading: MA: Addison-Wesley.

27. Spool, J., & Schroeder, W. (2001). *Testing web sites: Five users is nowhere near enough*. Available from: http://www.winwriters.com/download/chi01_spool.pdf

28. Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge, England: Cambridge University Press.

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

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

31.Vossen, P.H., & Maguire, M. (1998). *Guide to mapping requirements to user interface specifications. RESPECT D4.2*, EC Telematics Applications Programme, Project TE 2010. Available from: http://www.ejeisa.com/nectar/respect/4.2/

32. Woodward, B. (1998). *Evaluation methods in usability testing*. Available from: http://www.swt.edu~hd01/5326/projects/BWOODWARD.HTML