Analysis of Combinatorial User Effect in International Usability Tests

Effie Lai-Chong Law
Eidgenössische Technische Hochschule Zürich
Gloriastrasse 35, CH-8902, Zürich, Switzerland
law@tik.ee.ethz.ch

Ebba Thora Hvannberg
University of Iceland
Hjardarhaga 2-6, 107 Reykjavik, Iceland
ebba@hi.is

ABSTRACT
User effect in terms of influencing the validity and reliability of results derived from standard usability tests has been studied with different approaches during the last decade, but inconsistent findings were obtained. User effect is further complicated by other confounding variables. With the use of various computational models, we analyze the extent of user effect in a relatively complex arrangement of international usability tests in which four different European countries were involved. We explore five aspects of user effect, including optimality of sample size, evaluator effect, effect of heterogeneous subgroups, performance of task variants, and efficiency of problem discovery. Some implications for future research are drawn.

Categories and Subject Descriptors: H.5.2 [Information Interfaces and Presentation]: User Interface – Evaluation/methodology

General Terms: Experimentation, Human Factors, Measurement

Keywords: International usability test, user effect, evaluator effect, binomial model, Monte Carlo simulation

INTRODUCTION

Usability tests have been extensively applied in industry to evaluate a system's prototypes of different levels of fidelity. Usability tests, in which the thinking aloud technique [14] is typically applied, have thus become a de facto standard usability evaluation method. The primary goal of a usability test is to derive a list of usability problems (UPs) from evaluators' observations and analyses of users' verbal as well as non-verbal behavior. Improvement requests are proposed to systems developers for correcting the UPs thus identified. Nonetheless, usability tests are costly in terms of time and manpower required. To enable the incorporation of usability tests into a product's development lifecycle, strategies to minimize the costs involved in running them are deemed necessary. Amongst others, reducing the number of participants recruited for a usability test is a commonly deployed strategy. The potential risk of such a strategy is the loss of significant data - severe UPs, which substantially undermine real end-users' performance and thus a system's acceptability.

Individual differences are often regarded as a nuisance in psychological empirical studies, because they tend to threaten the generalizability of research findings. Hence, the number of participants employed for these studies is usually set to be large so as to mitigate the effects of the inherent sample heterogeneity with the help of appropriate statistical methods. However, this approach normally is not applied in usability tests, given the constraint of cost reduction. Practitioners are often confronted with the tradeoff between minimizing the number of participants and maximizing the scope of findings. The question "What is the optimal number of users to yield the best possible results from a usability test?" is especially challenging.

We define 'user effect' as: *The varied capacities of individual users as defined by their respective expertise and experiences to capture a subset of detectable usability problems of a system, given the particularities of the context of a usability test*. Here the context refers to a cluster of interrelated factors such as fidelity of prototype, design of task scenarios, physical settings, rapport with experimenters, etc. Clearly, user effect has substantial impacts on the reliability and validity of results of usability tests. This problem has been investigated by a group of researchers [2, 4, 11, 12, 15, 16, 17, 20, 22, 23, 25, 26, 27], though the number is dwarfed as compared with those engaged in studying other HCI issues. With the deployment of different mathematical models and analysis tools, these researchers have inferred some inconsistent and open claims. Besides, evaluator effect, which resembles user effect, can be observed in different usability evaluation methods such as user test and inspection (e.g., cognitive walkthrough) where no real users are involved [11].

Subsequently, we are going to address six issues germane to user effect with reference to the data we have garnered in our international usability tests, in which 17 users and 2 evaluators from four different European countries (Iceland, Slovenia, Switzerland, UK) were involved.

LITERATURE REVIEW AND RESEARCH QUESTIONS

In this section, we will investigate five aspects of user effect. We will first briefly delineate the previous works and then formulate six research questions (***R***).

Optimality of Sample Size

The binomial probability formula:

$$1 - (1-p)^n \qquad \textbf{(Formula 1)}$$

where n is the number of users or evaluators involved in a usability test and p is the probability of finding the average UP when running a single, average user (i.e., problem discovery/detection rate). Accordingly, five users are necessary to capture 80% of the known UPs of a system when p was around 0.3. Besides, additional users always increase the number of UPs, but with diminishing returns. Virzi [26] modeled the accumulation of UPs with increasing numbers of participants by Monte Carlo procedure and found that the resulting curve fitted well with that based on the binomial model. Some other studies supported the applicability of the binomial model to predict the proportion of UPs uncovered. However, the so-called "magic five" proposition was questioned [13]. Results of some field as well as experimental studies challenge the generalizability of this model to a variety of usability tests with diverse contexts [2]. In fact, the two basic assumptions underlying Formula 1 are violated: individual task scenarios are not independent events and individual problems have unequal likelihood to be identified. According to Herztum and Jacobsen [11] and Lewis [16], p can be much inflated when a sample size is small. To deal with this overestimation bias, the former have developed "any-two agreement" as an alternative measure to estimate the between-evaluator reliability, whereas the latter has demonstrated that taking the average of a normalization procedure and Good Turing discounting (i.e., hybrid adjustment approach) could best adjust p.

R1: *How do the three methods, Monte Carlo simulation, any-two agreement and the hybrid adjustment approach relate to each other and differ in terms of predicting the proportion of detectable usability problems?*

Evaluator Effect

Evaluator effect has increasingly attracted the attention of researchers [3, 9, 12, 24]. One of the salient issues addressed is the judgment of problem severity, which is primarily determined by evaluators' competences to assess problem impact accurately and consistently. Scales of varied granularity (e.g., major vs. minor or seven-point rating) and definitions of varied levels of abstraction (e.g., severity in terms of impact and frequency or a list of usability criteria) are employed for characterizing severity of UPs. Nonetheless, exercising judgment is highly subjective. Regardless of which scale or definition is employed, evaluators tend to apply it in a personalized and contextualized manner. This phenomenon can be well explained with the situated-constructivist approach [10]. Between-evaluator discrepancy is a rule rather than an exception, but within-evaluator discrepancy is not uncommon. The studies on evaluator effect focus on the former but neglect the latter. However, within-evaluator inconsistency can intensify evaluator effect to an appreciable extent, which in turn aggravates user effect.

R2: *Are between-evaluator discrepancy and within-evaluator discrepancy of comparable magnitude?*

Effects of Heterogeneous Subgroups

Whereas significant correlations between problem- discovery rate and problem-severity level were evident in some studies [19, 26], such results could not be verified in others [15]. Hence, we can say that not all severe problems have high p. The decoupling of these two parameters has a crucial implication for selecting sample size. Caulton [4] attempted to explain these inconsistent findings with the idea of heterogeneous subgroups. Specifically, he defined two types of UPs: *shared* UPs that can be detected by more than one group in the sample consisting of several subgroups, and *unique* UPs that can only be detected by one of these subgroups. By aggregating heterogeneous subgroups into a sample and treating them as a homogeneous one, the value of p of certain *unique* UPs can be diluted and the correlation between problem severity and problem frequency can be masked. Consequently, the power of a usability test will be assessed to be lower than it should be, where power is defined in terms of the number of users required to uncover a certain percentage of detectable UPs. Nonetheless, Caulton's thesis entails more support from empirical data.

R3: *To what extent does the presence of heterogeneous groups mask the correlation between problem discovery rate and problem severity level?*

R4: *How do the effects of different group characteristics in diluting the problem discovery rate differ?*

Performance of Task Variants

Selecting the core features rather than including all features of a system for testing may systematically favor one subgroup of participants, generating the heterogeneity effect described above [4]. Lewis [15] suggested that the likelihood of discovery of a specific problem could be increased if participants are required to perform the same task repeatedly or a task variant. In fact, repeating tasks can be a means to evaluate the learnability of a system, which is evident by the mitigation of the problems experienced in the initial attempt or by the reduction of time-on-task (cf. practice effect).

R5: *Does the performance of a task variant increase the likelihood of detecting specific problems?*

Efficiency of Problem Discovery

Efficiency is one of three canonical usability metrics. In accord with Common Industry Format, efficiency of a task is computed through dividing its unassisted completion rate by its mean time-on-task. However, the resulting value is not of any particular significance unless it is used as a benchmark for comparing similar products. Furthermore, in ISO/IEC 9216 Software Engineering – Product Quality Standard, metrics for different characteristics of usability are defined, but the problem discovery rate is not taken into account. In usability tests typically no time constraint is imposed on

performing a task. We assume that efficiency can alternatively be defined as the number of UPs that a participant can detect during the period of time when he or she is actively engaged in a certain task (i.e., time-on-task).

R6: *Is the efficiency of problem discovery a valid and objective criterion for assessing whether a participant is an experienced or a novice user?*

INTERNATIONAL USABILITY TESTS (IUT)

The system on which we performed IUT was a platform designed for enabling the exchange of online educational content among academic and industrial institutions. The user interface of this brokerage platform has been translated from its original English version into different European languages. The primary goal of IUT is threefold: ensuring the acceptability of the translation, identifying UPs, and assessing culture-dependent usage behavior. Four versions were tested: English, German, Icelandic, and Slovenian.

Design

Standard user test procedures [8] were adopted. IUT were conducted indigenously with local testers interacting with native participants in native language in the local context. The IUT Coordinator, a usability specialist, developed testing materials and tester guidelines to ensure the highest possible uniformity and quality of the tests. Local Testers were responsible for implementing the tests, recording the data, transcribing and translating thinking aloud protocols of participants. The involvement of Local Testers was essential, given the language barrier between the test designer and the participants [18]. The qualifications of Local Testers were being native speaker, knowledgeable in HCI and fluent in written English. All the raw data were sent to the IUT Coordinator for further processing.

Participants

The minimum number of participants per site was set to three, considering the limited resources available. Altogether 19 participants were involved: 4 English, 7 German, 5 Icelandic, and 3 Slovene native speakers. There were 6 researchers, 4 university professors, 3 teachers, 2 project managers, 2 administrators, 1 system developer, and 1 librarian. Their heterogeneous levels of competence in information technology and e-Learning could account for the diversity of usage behaviors observed.

Tasks

Each participant was asked to perform ten task scenarios covering the core functionalities of the platform, including applying for a user account, providing and offering learning resources, modifying different attributes of the learning resource provided, updating offers, searching and browsing the catalogue, and accessing selected learning resources.

Procedure

Participants were escorted into a testing room and seated at a desk with a computer system. They were asked to maintain a running commentary as they interacted with the system. They were asked to complete a pre-test and a post-test questionnaires, and an "after-scenario questionnaire" for each of the ten tasks. The test sessions were videotaped. The average time-on-task over ten tasks was 46.3 minutes.

Data Analysis

Quantitative and qualitative performance data were collected. The former included time-on-task, number of different errors, frequency of help sought, and instance of expressed frustration. The latter included the participants' thinking aloud protocols and the Local Testers' observations. The data of two participants (1 Icelandic and 1 English) were discarded, because they attempted only a small subset of the ten given tasks. Two evaluators, who were knowledgeable in usability evaluation methods, were involved in extracting UPs from the qualitative data collected. For each of the four testing sites, a separate list of UPs was prepared. Then, the four lists of UPs were merged together to produce a master list of non-overlapping UPs. Duplicate UPs were eliminated with the procedures similar to those employed by Connell and Hammond [5], and problem instances rather than problem types were counted. Each of the UPs in the master list was rated as severe, moderate or minor according to the conventional definitions [1].

RESULTS

In view of the limited space, only the findings related to the above research questions, of which the codes are quoted in parentheses next to the section headings, will be reported. The equations below are used for computing different parameters:

Detection rate p = Average of $|P_i| / |P_{all}|$ over all n users (**Eq. 1**), where P_i is the set of problems identified by *user* $_i$ and P_{all} is the set of problems identified collectively by all *n* users.

Any-two agreement = Average of $|P_i \cap P_j| / |P_i \cup P_j|$ over all ½ *n* (*n*-1) pairs of users (**Eq. 2**), where P_i and P_j are the sets of UPs identified by $user_i$ and $user_j$, and *n* is the number of users [11].

Hybrid adjustment:

adjp = ½ *[(estp-1/n)(1-1/n)] + ½[estp/(1+GTadj)]* (**Eq. 3**) where *adjp* is the adjusted estimate of *p* (*estp*) calculated from (Eq. 1), *n* is the sample size, and *(1-1/n)* is the lower limit of the "true" *p* and *GTadj* is the Good Turing adjustment to probability space, which is the proportion of the number of problems that occurred once divided by the number of different problems [16].

Descriptive Statistics

A total of 95 UPs were identified based on the data of 17 participants. Seven of them were caused by inadequate translation and excluded from the ensuing analyses to avoid the possible diluting effect mentioned in the foregoing discussion. Table 1 shows the number of UPs identified in each of the four language versions. The column 'unique UPs'

indicates the number of UPs that were identified by only one specific group, but basically could have been detected by other groups if more participants were involved.

	Total UPs	Unique* UPs	Shared UPs
English (3)	21	6	15
German (7)	68	36	32
Icelandic (4)	25	7	18
Slovenian (3)	21	6	15

Table 1. Distribution of all UPs over four language versions.

The mean problem detection rate p of 88 problems over 17 participants was **0.14** (SD = 0.07), ranging from 0.05 (4 problems) to 0.32 (28 problems). The total number of unique problems, which were identified by a single user, was 41. There was only one problem commonly identified by 9 participants (see Figure 1).

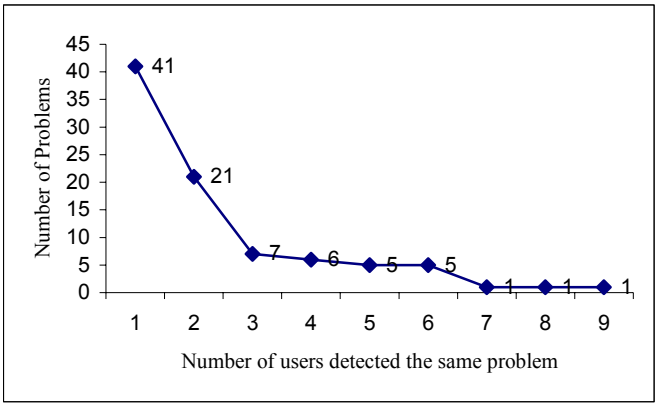

Figure 1. Prevalence of usability problems

Detection rate, Any-two agreement and Hybrid adjustment (*R1*)

A Monte Carlo procedure was applied to the data of UPs to derive the general form of the curve relating the proportion of UPs identified to the number of participants involved in the usability test [6]. A computer program was developed to generate 500 permutations of the participant order and to calculate the mean number of unique problems identified at each sample size (1-17). The resultant curve is shown in Figure 2. Besides, a curve based on binomial model (Formula 1) is plotted with p being equal to 0.14 - the mean probability of problem detection in the current sample. The two curves fit notably well.

Language is the single factor distinguishing the four versions of the system (i.e., minimal localization). We computed within-group problem discovery rates and compared them with the corresponding rates based on the pooled 88 UPs (see Table 2). The p_{within}/ $p_{overall}$ ratios (i.e., inflation rate) range from 1.39 (German group) to 5.0 (English group). We applied Monte Carlo (MC) simulation to each of the four groups and plotted the curves, which were overlaid with the curves constructed based on the binomial model (BM) with the respective detection rates. Figure 3 displays the results. Generally speaking, the fitness of the curve is inversely proportional to the group size. The variations in fitness can be attributed to the different inflation rates.

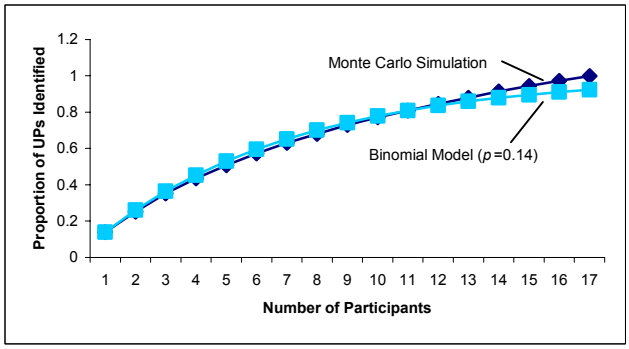

Figure 2. Monte Carlo simulation vs. binomial model for all participants.

	No. of UPs *within* group	p_{within}	$p_{overall}$
English (3)	21	0.40	0.08
German (7)	71*	0.25	0.18
Icelandic (4)	29*	0.34	0.11
Slovenian (3)	21	0.56	0.12

Table 2. Within-group problem detection rates.
(NB: * UPs caused by translation were included)

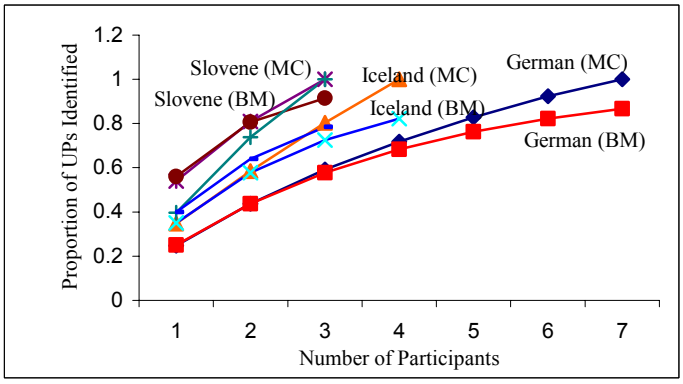

Figure 3. Monte Carlo (MC) simulation vs. binomial model (BM) for four different groups.
(NB: The two unlabelled curves are English; the lower one is BM)

Furthermore, we calculated any-two agreement (Eq. 2) and the hybrid adjustment of p (Eq. 3) for the sample as a whole and also for each of the four groups (Table 3). Among the four subgroups, the English has the lowest value for any-two agreement. In other words, the participants of this group tended to identify unique problems. Except for the English group, the higher the value of any-two agreement, the lower the number of participants was. Similarly, the higher the value of non-adjusted p, the lower the number of participants was. Generally speaking, any-two agreement and non-adjusted p should be significantly correlated, because the more UPs two users identified independently, the higher the probability that they will identify the same ones. Hence, while p and any-two agreement cannot be equal, because they are calculated based on different mathematical models, there should be significant correlations between them. Note that the discrepancies between any-two agreements and non-adjusted ps are substantially larger than those between any-

two agreements and hybrid-adjusted *p*s. The Pearson correlation between the former two values is 0.683, whereas the Pearson correlation between the latter two is 0.916.

(n)	All (17)	English (3)	German (7)	Iceland (4)	Slovenian (3)
Any-two Agreement	0.09	0.09	0.16	0.18	0.32
Hybrid-adjusted *p*	0.09	0.13	0.13	0.13	0.27
Non-adjusted *p*	0.14	0.40	0.25	0.34	0.56

Table 3: Any-two agreement and hybrid adjustment.
(Note: non-adjusted *p*s are equal to p_{within} of Table 2)

Within- and Between-Evaluator Consistency (*R2*)

The UP extraction task was particularly challenging in the current case because of the language barrier. First, the two evaluators, E1 and E2, read through the translated verbal protocols and the Local Testers' observation reports, and then examined the videotapes. The two evaluators underwent two rounds of UP extraction. During the first round we initially worked on the German group independently, but found that the agreement was disappointingly low. Hence, for the sake of mutually understanding each other's extraction methods, we collaboratively worked on the data of some selected participants. Then, we analyzed the remaining data independently. Table 4 shows the extraction results.

	English	German	Icelandic	Slovenian
E1	13	58	15	18
E2	20	64	25	21

Table 4: Numbers of UPs identified by the two evaluators.

E1 consistently identified fewer problems than E2 did. It could be attributed to the fact that E2 was a more experienced evaluator and was more familiar with the system tested. Discrepancies were negotiated. Some UPs were further split and some were collapsed. After finalizing the four UP lists, E2 merged them into a master list with 87 non-overlapping UPs. E1 and E2 then judged the severity of individual UPs independently. The Kappa measure for the inter-rater reliability in this judgment exercise was 0.64.

About two months later, E1 and E2 repeated the extraction exercise to check the reliability. Some UPs that had not been identified in the previous attempt were uncovered. Interestingly enough, some UPs that had been identified previously were not "re-uncovered". E1 and E2 repeated the severity rating exercise of the updated master list of UPs independently. Any-two agreement for the problem extraction (within-evaluator) for E1 and E2 were 0.79 and 0.83, respectively, and any-two agreement for the problem extraction (between-evaluator) was 0.71. The Kappa measure for the severity ratings for the overlapping UPs between the two attempts (within-evaluator) for E1 and E2 were 0.74 and 0.82, whereas the Kappa measure for the severity ratings for the updated master list of UPs (between-evaluator) was 0.69. Though the extents of within-evaluator disagreements were lower than those of between-evaluator disagreements, the fact that individual evaluators cannot reliably extract UPs or judge the severity of UPs is a concern to be addressed. Nevertheless, the two evaluators discussed the disagreements and drew the consensus.

Problem Severity and Subgroup Effect (*R3 & R4*)

88 UPs were categorized into three severity levels and the respective problem detection rate for each of the three UP severity groups was computed (Table 5).

	Minor	Moderate	Severe
Frequency	25	40	23
p	0.1	0.13	0.21

Table 5. Distribution of problem types.

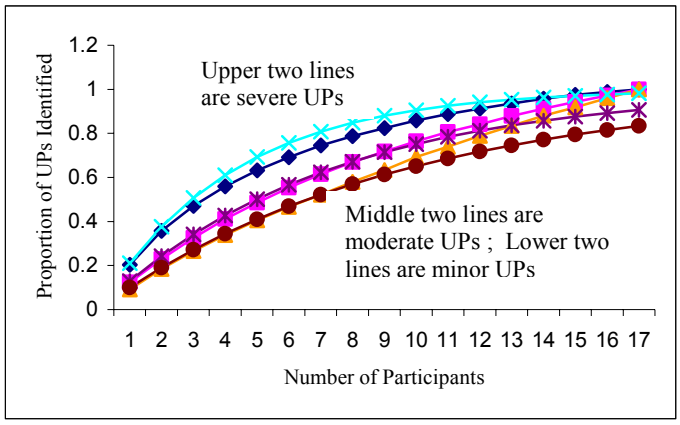

Figure 4. Monte Carlo simulation vs. binomial model for three types of UPs.

Five out of 23 severe problems were identified by only a single participant. The problem detection rate for the severe UPs was significantly higher than that of the less severe ones (i.e., moderate plus minor) (t=3.4794; p=0.0008). We applied Monte Carlo simulation to model the increment of the proportion of UPs detected as the function of the increasing number of participant for all the three types of UPs. The resulting curves (MC) were mapped against the curves derived from the binomial model (BM) with the respective *p*s. The BM curve for the severe problems is consistently above its MC counterpart till the number of participant is 14. This tendency of overestimation was explained as a Jensen's Inequality artifact – the inherent feature of the mathematical modeling adopted [15, 26]. The BM curves for the moderate and the minor, however, do not reveal such a pattern. It seems that the fitness of the two types of curves is determined by the value of p. The lower the p, the lower the fitness will be.

Correlation between the problem severity and problem detection rate was computed. In contrast to Virzi's [26] but consistent with Lewis' [15] findings, the resulting correlation (r = 0.0963, df = 381, p=0.03) is *not* significant.
We then applied Caulton's [4] approach to investigate whether the insignificance can be attributed to the effects of heterogeneous subgroups. First, we assumed the cultural background as the between-group variable. Nine of 23 severe problems were uniquely identified by one of the four cultural subgroups and their corresponding *p*s were obviously "diluted". For instance, for UP8 (i.e., the system accepted the request for retrieving a forgotten password even when the

username entered was invalid), only one of three English users identified it (p_{within} = 0.33). When averaging this p_{within} over 17 participants, the p_{all} was shrunk to 0.06. Similarly, the *p*s of three other severe problems were diluted by the factor of 5. Table 6 displays the average severe problem detection rates of the four groups, which do not indicate any diluting effect. Certainly, when the size of individual subgroup is small, the diluting effect will be exaggerated.

p(all)	*p*(English)	*p*(German)	*p*(Iceland)	*p*(Slovenian)
0.21	0.17	0.25	0.14	0.22

Table 6. Severe problem detection rates of four subgroups.

Further, we assumed the self-reported expertise as another between-group variable. Two groups were thus identified – experienced and novice system users. The classification criterion was based on the average rate (5-point scale) that individual participants ascribed to their own competence in information technology, experience in operating database systems, and experience in e-learning. Those who scored three or less were categorized as novice. Five of the 23 severe problems were identified only by experienced users and another five severe problems were identified only by novice users. On average, the diluting effect observed in two-level expertise groups was apparently lower than that observed in the four cultural groups, with the highest ratio of $p_{experienced}/p_{all}$ and p_{novice}/p_{all} being two. This finding is consistent with Caulton's claim that the higher the number of subgroups, the more serious the diluting effect will be.

Performance of Task Variants (*R5*)

The brokerage platform tested supports the exchange of two major types of learning resources, namely educational material (EM) and educational activity (EA). Provisions of EM and EA involve basically describing a set of common core metadata attributes and some additional ones specific to EA. The participants were required to provide EM (Task 2) and EA (Task 4). The rationales of including both tasks were to investigate whether participants would have problems in describing specific metadata attributes and to evaluate the learnability of these tasks. To verify the claim that performing task variants can enhance problem discovery [15], we adopted a somewhat simplified case study approach and examined the data of the seven German participants. Table 7 shows the results of analyses. P3 identified ten additional UPs with two of them being severe. P4's data showed a similar pattern. P5 and P7 experienced one UP that was already found in Task 2.

The average time-on-task (in minutes) of Task 2 over the seven participants (*M*=14.74, *SD*=4.3) was higher than that of Task 4 (*M*=9.95, *SD*=2.6), though the participants were required to describe additional attributes for Task 4, and the difference was statistically significant (*t*=2.1344; *df*=6; *p*=0.0767). This result suggests the learnability of the task.

Participant	P1	P2	P3	P4	P5	P6	P7
Task 2	4	8	7	2	4	6	5
Task 4							
-Add	**1**	**0**	**10**	**8**	**1**	**0**	**2**
-Spec	3	0	2	1	1	2	0
-Dup	0	0	0	0	1*	0	1*
-Type	1Mo	--	2S 3Mo 5 Mi	2S 4Mo 2Mi	1Mi *1Mi		1S *1M

Table 7. Results of performing task variants
(NB: Add: additional UP; Spec = UP specific to Task 4 attributes; Dup = Duplicate UP; Type = (S)evere)/ (Mo)derate/ (Mi)ld)

Correlation between Time-on-Task & Number of UP (*R6*)

To investigate the relationship between time-on-task and number of UPs identified, we examined the data for the two most complex and problematic tasks: Task 2 (Providing new educational material) and Task 4 (Providing new educational activity). All the participants experienced UPs in Task 2 and all except one participant experienced UPs in Task 4. The ranges of the total number of UPs identified are 1 to 8 and 1 to 12 for Task 2 and Task 4, respectively. The Spearman correlation between the numbers of UPs and the times-on-task of both tasks for all the participants is insignificant (r = 0.0356).

When time-on-task of a specific task is longer than the corresponding benchmarked value, it typically implies the existence of UP. But the current results seem to refute this assumption. In fact, a number of participants, when confronted with a UP that could not be circumvented, tended to repeat the same action sequence, resulting in longer time-on-task. In some cases, the participants, when confronted with UPs, gave up prematurely without attempting to work around them. Some participants, who were so motivated to identify UPs, tended to check every detail of the system, resulting in longer time-on-task. Table 8 shows the average numbers of UPs per minute of time-on-task over all the participants for Task 2 and Task 4.

	Average	*SD*	Min.	Max.
Task 2	0.42	0.28	0.15	1.23
Task 4	0.36	0.31	0.00	1.22

Table 8. Average number of UPs per minute for two tasks

We computed the correlations between individual participants' numbers of UPs per minute with their self-reported expertise (see above). The Spearman correlations for Task 2 (r = 0.0677) and Task 4 (r = 0.0712) are insignificant. The findings suggest that time-on-task is an elusive variable, because the value is determined by a number of intertwined factors, such as expertise of participants, especially problem-solving behavior, and motivation of participants.

GENERAL DISCUSSION

The following discussion is indexed by the codes of the research questions posed above.

R1: Our results clearly show that the so-called "magic five" assumption cannot be held. To obtain 80% of the detectable UPs of the system tested, 11 participants were required

(Figure 2), given the relatively low problem discovery rate (*p*) 0.14. However, this *p*, calculated with the use of the conventional model, is probably inaccurate, because the basic assumptions underlying the model are violated. Can the problem detection rate be better estimated by the parameter any-two agreement, which has been originally developed to counteract the overestimation bias inherent in *p* and presumably gives a more accurate value of inter-evaluator reliability? Our findings (Table 3) indicated that any-two agreement tended to underestimate *p*. This observation was further verified by the results obtained from our simulation program, in which the values of three parameters (number of participants, number of UPs, and 'true' *p*) were input. Suppose that the true *p* is 0.5, two users or evaluators are involved, and altogether *Z* problems are identified. $|P_i \cap P_j|$ will be equal to ¼ *Z* problems and $|P_i \cup P_j|$ will be equal to ¾ *Z* problems on the average. The resulting any-two agreement will be 1/3, i.e., underestimating the true *p* by the factor of 2/3. The relationship between any-two agreement (*A*) and *p* (estimated from the sample) can be approximately represented as:

$$p = (2 * A) / (1 + A) \qquad \textbf{(Eq. 4)}$$

In summary, while the estimated *p* computed with Eq.1 tends to overestimate the true *p*, any-two agreement (Eq. 2) tends to underestimate it. Furthermore, as demonstrated by Lewis [16] with the use of Monte Carlo simulation, the hybrid adjustment (Eq.3) can lead to a good estimate of true *p* when the number of users is less than or equal to ten. Our findings (Table 3) show that the non-adjusted *ps* are consistently higher than the corresponding hybrid-adjusted *ps* by a factor of two or even three. Future research should be invested in verifying the applicability of this approach to the sample size much larger than ten.

R2: Within-evaluator inconsistency, as evident by our findings, is a hitherto neglected issue that needs to be addressed. Between-evaluator agreement is difficult to reach if evaluators are not consistent in applying their strategies for extracting problem and rating problem severity. We propose that individual evaluators check the reliability of their own ratings with two rounds of evaluations that are separated by at least one-week gap. The between-evaluator agreement index should be adjusted by taking the within-evaluator agreement into account, resulting in so-called combinatorial evaluator-agreement index ($A_{\text{combinatorial}}$) (see Eq. 5):

$$A_{combinatorial} = A_{between} * A_{within} \qquad \textbf{(Eq. 5)}$$

where A_{between} and A_{within} can be computed according to Eq. 2 or a better method to be identified. Note, however, Eq. 5 is a simplified model to represent the interactive effect of between- and within-evaluator reliability. Future work is definitely required to improve the equation. Another evaluator effect, which is amplified in the context of international usability tests (IUT), is the two-tiered evaluation. In our studies, the four Local Testers performed the low-level data collection while the two evaluators performed the high-level problem extraction and severity rating. Inevitably, Local Testers might bias their observations or even translations of thinking aloud protocols because of their personal experiences and expertise. Consequently, the biased data would likely undermine the validity of the results. This so-called Local Tester or experimenter effect, to our knowledge, is not yet addressed systematically in the literature. UTs are costly, but IUTs cost even more. Expert evaluators are expensive resources. In principle, more accurate results will be produced if native evaluators are employed to extract UPs from the data presented in native language. However, the costs involved may outweigh the benefits. The issue of return on investment (ROI) is tricky and complex [15] and can probably be explored with in-depth case studies.

R3 & R4: There is no doubt that the heterogeneity of subgroups in a sample will dilute the problem discovery rate. As indicated by our results, not only the *p*s of severe problems but also those of the moderate and minor ones were shrunk due to the diluting effect, and the respective shrinkages were of similar degree. In fact, quite a number of minor and moderate UPs were detected only by the German group. Given that the numbers of participants were different in the four cultural subgroups, the diluting effects might have been exaggerated.

The problem discovery rate (*p*) of the severe problems is significantly higher than that of the less severe problems. But the absolute value of *p* for the severe problems is *not* particularly high. As illustrated in Figure 4, between nine and ten participants were required to uncover 80% of the severe problems, whereas 15 participants were required to uncover 80% of the minor problems. In summary, it would be somewhat risky if we recruited only five participants in our IUT, because only 68% of the severe problems would be detected. Furthermore, there was no significant correlation between problem detection rate and problem severity level for the sample as a whole. Nor for the German group when we computed the correlation with somewhat "un-diluted" *p*s. Clearly, the decoupling of these two parameters is one of the reasons for usability practitioners to recruit more participants for usability tests. For IUTs, the recommended number of participants is six [7], however, the choice seems arbitrary because no explicit rationale is given.

R5: To a certain extent, our results supported Lewis' [15] claim that the opportunity to perform a task variant will enhance the likelihood of detecting additional UPs. As evident in our case analyses of seven participants performing two similar tasks, altogether 22 additional UPs were identified with 5 of them being severe problems. Intuitively, according to the practice effect, the persistence of UPs was low (only two instances across the seven participants) because the participants had found some means to work around them. These findings seem to suggest that it is desirable to include task variants in usability tests. However, as pointed out by Lewis there is a tradeoff between the increased duration of test sessions and the increase in the number of UPs detected.

R6: The time-on-task is such a multifarious variable that great caution should be taken when it is used to indicate efficiency, be it computed according to the standards such as Common Industry Format or related to the number of UPs detected. Nonetheless, we assume that investigating the relationship between the time-on-task that evaluators invest in inspecting a system (e.g., heuristic evaluation) and the number of UPs that they thus identify will yield significant results, considering that they normally do not attempt to solve UPs uncovered and thus do not unduly lengthen the time-on-inspection task.

CONCLUDING REMARK

Although the six research questions we addressed in this paper cannot be perfectly answered with our data, they are definitely the issues that necessitate attention and efforts of usability researchers and practitioners. Given that a host of interrelated factors can potentially affect the reliability and validity of results of usability tests, we need more robust mathematical and advanced statistical models to help us understand the related issues and to enable us make more accurate predictions. Furthermore, we re-emphasize our recommendation proposed elsewhere that when a sufficient number of systematic, well-designed and professionally performed empirical works on usability tests are available, *meta-analysis* can be conducted on them to infer a clear, holistic, and more conclusive picture about the issues addressed in this paper.

ACKNOWLEDGEMENTS

The authors would like to thank Thomas Erlebach for implementing Monte Carlo simulations and discussing some of the results, and are also grateful to all the participants of the International Usability Tests.

REFERENCES

1. Artim. J. M. (2003). Usability problem severity ratings. Access at: http://www.primaryview.org/CommonDefinitions/
2. Bailey, R.W. (2000). *Improving usability in America's voting systems*. Access at: http://www.humanfactors.com/library/nov00.asp.
3. Catani, M.B., & Biers, D.W. (1998). Usability evaluation and prototype fidelity. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, pp.1331-5.
4. Caulton, D.A. (2001). Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology, 20*(1), 1-7.
5. Connell, I. W., & Hammond, N.V. (1999). Comparing usability evaluation principles with heuristics. In *Proceedings of INTERACT'99*.
6. Diaconis, P., & Efron, B. (1983). Computer intensive methods in statistics. *Scientific American, 248*, 116-130.
7. Dialogdesign – international usability testing. Access at: http://www.dialogdesign.dk/internationaltest.html
8. Dumas, J.S., & Redish, J.C. (1999). *A practical guide to usability testing* (rev. ed.). Exeter: Intellect.
9. Fu, L., Salvendy, G., & Turley, L. (1998). Who finds what in usability evaluation. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting.*
10. Greeno, J.G., Collins, A.M., & Resnick, L.B. (1996). Cognition and learning. In D.C. Berliner & R.C. Calfee (Eds.), *Handbook of Educational Psychology*.
11. Hertzum, M., & Jacobsen, N.E. (2001). The evaluator effect *International Journal of Human Computer Interaction, 13*(4), 421-444.
12. Jacobsen, N.E., Hertzum, M., & John, B.E. (1998). The evaluator effect in usability studies. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting.*
13. Law, L.-C., & Hvannberg, E.T. (2002). Complementarity and convergence of usability tests and heuristics evaluation. In *Proceedings of NordiCHI 2002.*
14. Lewis, C. (1982). Using the thinking aloud method in cognitive interface design. Yorktown Heights, NY: IBM Thomas J. Watson Research Center. (RC 9296#40713).
15. Lewis, J.R. (1994). Sample sizes for usability studies: Additional considerations. Human Factors, 36(2), 368-378.
16. Lewis, J.R. (2001). Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human-Computer Interaction, 13*(4), 445-480.
17. Molich, J., et al. (1998). Comparative evaluation of usability tests. In *Proceedings of UPA 98*.
18. Murphy, J. (2001). Modeling "Designer-Tester-Subject" relationships in international usability testing. In *Proceedings of IWIPS 2001*, pp. 33-44.
19. Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In P. Bauersfeld, J. Bennett & G. Lynch (Eds.), *Proceedings of CHI'92*.
20. Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. International Journal of Human-Computer Studies, 41, 385-397.
21. Nielsen, J. (2000). "Why you only need to test with 5 users", Available at http://www.useit.com/alertbox/20000319.html.
22. Nielsen, J., & Landauer, T.K. (1993). A mathematical model of the finding of usability problem. In *Proceedings of InterCHI'93*.
23. Spool, J., & Schroeder, W. (2001). Testing websites: Five users is nowhere near enough. In *Proceedings of ACM CHI Conference on Human Factors in Computing*.
24. Vermeeren, A., van Kesteren, I. & Bekker, M. (2003) Managing the 'Evaluator Effect' in User Testing, In *Proceedings of Interact 2003*.
25. Virzi, R.A. (1990). Streamlining the design process: Running fewer subjects. In *Proceedings of the Human Factors Society 34th Annual Meeting.*
26. Virzi, R.A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors.*
27. Woolyrch, A., & Cockton, C. (2001). Why and when five test users aren't enough. In Proceedings of IHM-HCI 2001 Conference (Vol. 2).