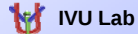


VALUTAZIONE SPERIMENTALE DI USABILITA' DI SISTEMI INTERATTIVI

Rosa Lanzilotti



IVU Lab

Dipartimento di Informatica

Università di Bari, via Orabona, 4 – 70125 Bari

lanzilotti@di.uniba.it



Sommario

- Metodi di valutazione dell'usabilità
- Metodi user-based
- Esperimento controllato
- Esempio di studio sul campo
 - valutazione di un sistema di m-learning
- Esempio di esperimento controllato
 - validazione di una tecnica di ispezione

Metodi di valutazione dell'usabilità

- Ispezioni
 - valutatori esperti ispezionano l'interfaccia utente con lo scopo di trovare possibili problemi di usabilità e fornire giudizi basati sulla loro conoscenza
- User-based
 - le proprietà dell'usabilità vengono valutate osservando come il sistema, o un prototipo del sistema, è utilizzato da utenti reali, o loro rappresentanti che svolgono compiti veri

Stili di valutazione

- Studi sul campo
 - ↑ consentono di osservare interazioni tra sistema e utente che si sarebbero perse in un laboratorio di usabilità
 - ↓ alti livelli di rumore ambientale rendono difficile l'osservazione sul campo
- Studi di laboratorio
 - ↑ offrono attrezzatura sofisticata che non può essere riprodotta nell'ambiente di lavoro (p.e. computer potenti, specchi bilaterali, ecc.)
 - ↑ manipolano il contesto al fine di individuare problemi o procedure meno usate
 - ↓ non sempre simulano situazioni del mondo reale (p.e. tecnologie mobili)

Scelta del metodo di valutazione

■ Fase del ciclo di sviluppo del sistema

- progetto vs implementazione

■ Stile di valutazione

- laboratorio di usabilità vs ambiente di lavoro

■ Tipo di misure

- misure qualitative vs misure quantitative

■ Risorse richieste

Ispezioni

- Economici e semplici da organizzare

- Soggettivi avendo una forte dipendenza dall'esperienza del valutatore

- Difficili da standardizzare

Esempi: valutazione euristica, cognitive walkthrough, ispezione con linee guida

Metodi di Ispezione

■ Valutazione Euristica

- esame delle singole schermate da parte dell'esperto sulla base di principi e linee guida (euristiche), eseguendo alcuni task rappresentativi

■ Cognitive Walkthrough

- focalizzato sulla valutazione di un progetto rispetto alla facilità di apprendimento
- il valutatore formula un piano per eseguire un task ed esplora il sistema, ricercando opportune azioni per il completamento del task

■ Ispezione formale di Usabilità

- ingegneri del software valutano il sistema con metodi simili a quelli di ispezione del codice

■ Ispezioni con Linee guida

- l'esperto ispeziona l'interfaccia per valutare se è conforme alla lista di linee guida dell'usabilità che ha disponibile

Metodi user-based

- Molto affidabili (se fatti bene)

- Difficile selezionare un campione “veramente” rappresentativo degli utenti reali

- Difficile far eseguire agli utenti task complessi dopo un breve training

- Dispendiosi

Metodi user-based

- Metodi basati sull'osservazione
 - osservazione diretta
 - uno o più valutatori osservano gli utenti mentre svolgono i loro compiti nel loro ambiente di lavoro
 - thinking aloud
 - il valutatore chiede all'utente di pensare ad alta voce quando utilizza il sistema
- Tecniche di survey
 - interviste
 - gli sperimentatori raccolgono esperienze, opinioni e motivazioni sul comportamento dell'utente
 - questionari
 - personale anche inesperto può distribuire i questionari per raccogliere informazioni sulla situazione attuale e per avere risposte specifiche
- Metodi sperimentali
 - esperimenti controllati
 - verificano in che modo modifiche di alcuni parametri del sistema si ripercuotono sull'usabilità del sistema

Esperimento controllato

E' uno dei più rigorosi metodi di valutazione

- Scopo
 - fornire evidenza empirica ad un'ipotesi
 - si effettua quando si intende valutare come un cambiamento nel progetto influisce sull'usabilità del sistema, verificando relazioni di causa-effetto mediante il controllo delle variabili sperimentali
- Disegno Sperimentale
 - scelta dei partecipanti (particolarmente importante!)
 - scelta delle variabili
 - scelta delle ipotesi
 - scelta del metodo sperimentale
 - analisi dei dati

Tipologie di metodi sperimentali...

- Metodo naturalistico (flessibile)
 - Si osserva il comportamento dei partecipanti nel loro ambiente naturale
 - Non ci sono ipotesi forti che richiedono un particolare insieme di procedure osservative
- Case-study (approccio meno flessibile)
 - Si limita l'ambiente in cui il partecipante si muove e si può intervenire a piccoli passi
 - Si osservano le risposte del partecipante ad alcune domande che possono essere poste

...Tipologie di metodi sperimentali

- Correlational research
 - Le procedure di misurazione devono essere attentamente definite e seguite precisamente
 - Può variare da una ricerca naturalistica ad una altamente limitata come una ricerca fatta in laboratorio
- Differential research
 - Si confrontano due o più gruppi di partecipanti
 - Le variabili sono misurate esattamente nello stesso modo in tutti i gruppi
 - I gruppi si formano sulla base di una variabile preesistente che non è sotto il controllo del ricercatore (sesso, IQ, classe socio-economica, ecc.)
- Experimental research
 - Simile alla differential research tranne che i partecipanti sono assegnati in modo casuale ai gruppi o alla condizione sperimentale

Scelta dei partecipanti

- Devono corrispondere alla popolazione di utenti finali che utilizzano l'applicazione da testare
 - l'ideale è coinvolgere utenti reali
- Il numero di partecipanti deve essere:
 - sufficientemente rappresentativo della popolazione di utenti finali
 - adeguato ai metodi statistici scelti

N.B.: si devono avere almeno 10 utenti per condizione

Variabili

- Le ipotesi vengono testate manipolando e misurando variabili sotto condizioni controllate
- Tipi principali di variabili:
 - **indipendenti**: caratteristiche dell'esperimento che vengono manipolate per creare condizioni diverse da confrontare
Esempio: numero di voci in un menu
 - **livello**: ogni valore della var. indep. che viene usato in un esperimento
Esempio: tre menu con 5, 7 o 9 voci => 3 livelli
 - **dipendenti**: variabili che possono essere misurate nell'esperimento
Esempio: velocità dell'utente nel selezionare una certa voce del menu
 - variabili dipendenti più comuni:
 - tempo di esecuzione di un task
 - numero di errori commessi

Ipotesi...

- E' una previsione sul risultato dell'esperimento
- Formulata in termini di variabili indipendenti e dipendenti
 - una variazione nella variabile indipendente dovrebbe causare una variazione nella variabile dipendente
Esempio: un incremento del numero di voci in un menu dovrebbe causare un rallentamento nella velocità di selezione di una voce in esso contenuta
- Scopo dell'esperimento è dimostrare che questa previsione è corretta

...Ipotesi

- Ipotesi nulla H_0
 - non c'è differenza nella variabile dipendente tra i livelli della variabile indipendente
- Ipotesi alternativa H_1
 - c'è dipendenza tra variabili indipendenti e dipendenti: quindi, le variazioni nelle variabili indipendenti hanno effetti sulle variabili dipendenti
- Il ragionamento statistico è conservativo:
 - si assume H_0 vera
 - se i risultati sono statisticamente significativi si rifiuta H_0

Metodo sperimentale

- Condizioni sperimentali:
 - **Condizione sperimentale:** è quella in cui le variabili sono state manipolate
 - **condizione di controllo:** è identica alla precedente, ma non c'è alcuna manipolazione sulle variabili
 - È necessaria per stabilire che è la manipolazione ad essere responsabile delle differenze misurate
- Between-groups
 - ogni partecipante viene assegnato ad una sola delle condizioni sperimentali
- Within-groups
 - i partecipanti sono esposti a tutte le condizioni sperimentali
 - per evitare effetti di apprendimento è necessario controbilanciare l'ordine secondo cui i task sono effettuati dai partecipanti
 - servono meno utenti rispetto al between-groups
 - minor probabilità di effetti indesiderati dovuti alla scelta dei partecipanti
- Mixed design

Analisi statistiche

- Statistica descrittiva: riassume, semplifica e descrive un ampio numero di misure
 - media, mediana, frequenza, varianza
- Statistica inferenziale: aiuta ad interpretare quello che i dati vogliono dire
 - valuta la differenza tra le medie dei gruppi: t-test, analisi della varianza (ANOVA), ecc.

Analisi statistiche

- Test parametrici se i dati seguono una distribuzione normale
- Test non parametrici con distribuzioni non normali
- La scelta del test statistico dipende da:
 - Ipotesi
 - Numero di variabili indipendenti
 - Livello di misura di ogni variabile dipendente

Livello di misura dei dati

- Scala nominale
 - Livello più basso di misura
 - Es: il luogo di nascita, l'affiliazione politica, la categoria diagnostica, sesso
 - I dati misurati con scale nominali vengono chiamati *dati nominali*
- Scala ordinale
 - Misurano una variabile in ordine di grandezza
 - Es: classe socio-economica (bassa, media, alta)
 - I dati misurati con scale ordinali vengono chiamati *dati ordinati*
- Scala ad intervalli
 - Hanno la proprietà delle scale ordinali a intervalli regolari tra valori consecutivi su una scala
 - Scala ad intervalli non ha il valore 0
 - Es: temperatura (lo 0 non indica assenza di calore)
- Scala proporzionale
 - Hanno tutte le proprietà delle scale precedenti con lo 0
 - Es: peso, distanza, tempo, volume, numero di risposte
 - I dati misurati con scale ad intervalli e scale proporzionali vengono chiamati *score data*

Il modello Decision-Tree

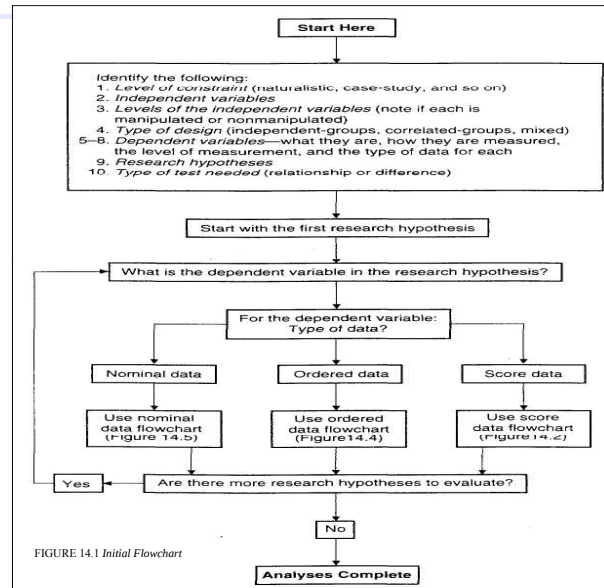


FIGURE 14.1 Initial Flowchart

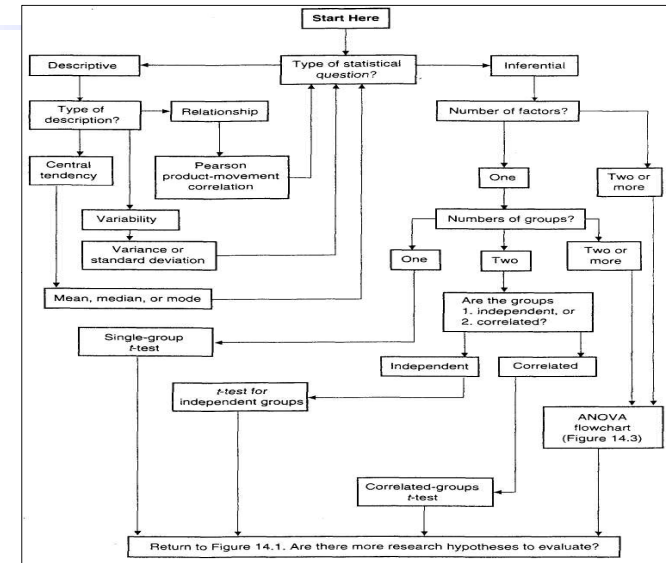
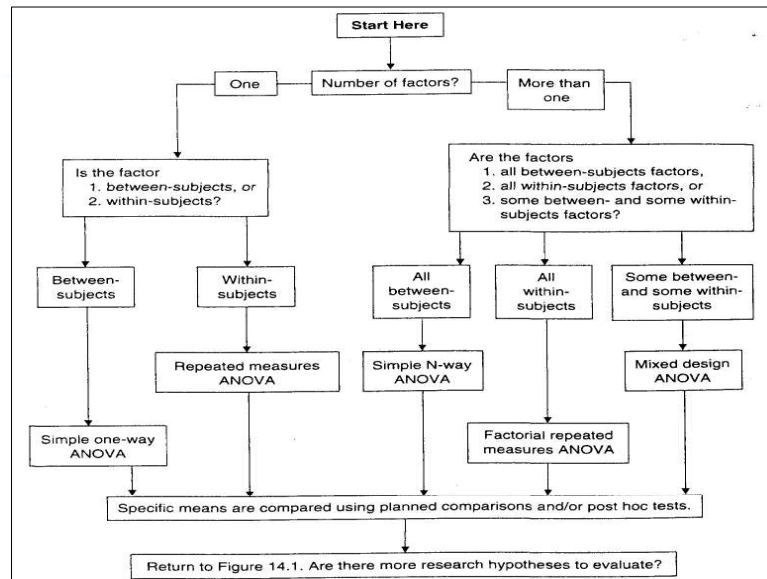
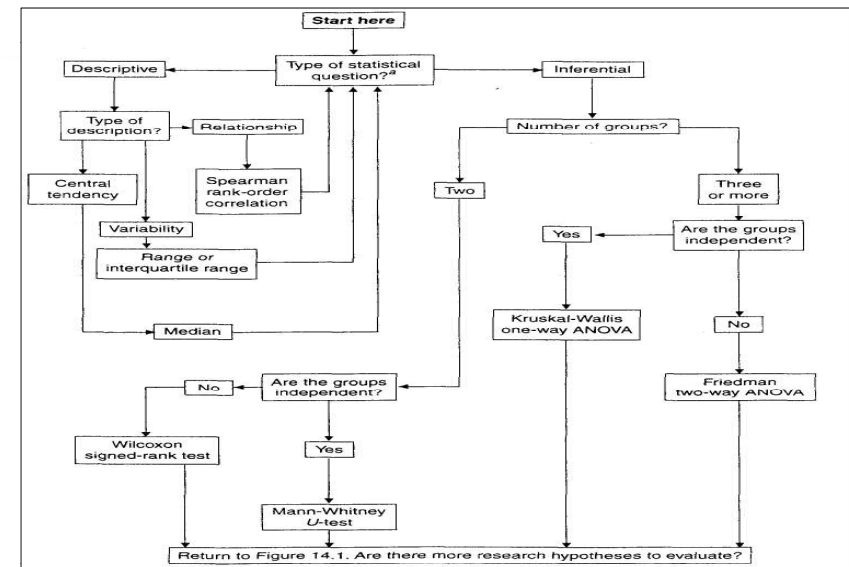


FIGURE 14.2 Score Data Flowchart



Note: Most formulas for the various ANOVAs and related statistical tests can be found in Appendix C. For additional computational procedures, consult an ad

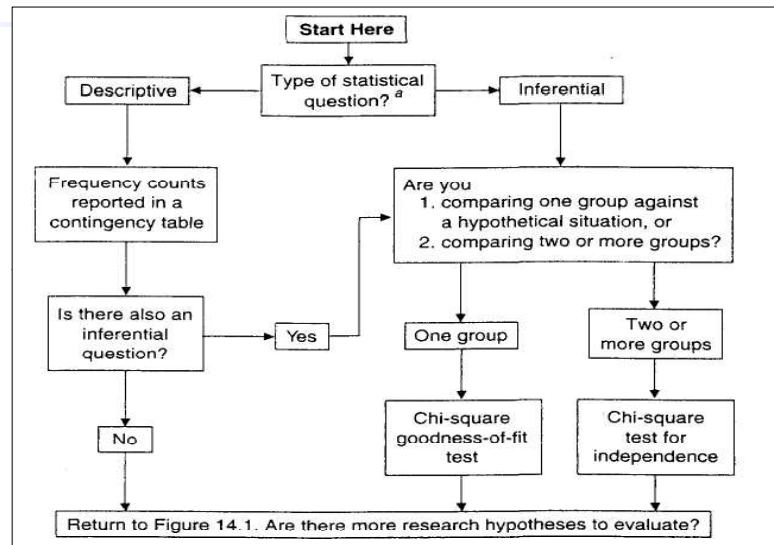
FIGURE 14.3 ANOVA Flowchart



* It is best to start with descriptive statistics. Note: Computational formulas for the statistical procedures described here can be found in Siegel (1956).

Riferimenti

- A. Dix, J. Finlay, G. Abowd, R. Beale, Interazione uomo-macchina, McGraw-Hill, 2004
- A. M. Graziano and M. L. Raulin, Research methods, a process of inquiry, Pearson, 2005



^aIt is best to start with descriptive statistics.

FIGURE 14.5 Nominai Data Flowchart

Pattern-based inspection

- It is based on the use of evaluation patterns
 - Evaluation patterns are descriptions of what an evaluator has to do when inspecting an application
 - Evaluation patterns describe tasks that an inspector should perform
 - Evaluation patterns provide guidance for inspectors
 - what should they look for?
 - where should they start from?
 - which application objects are more critical and worth to invest more time?
 - which actions should they perform?

VALIDAZIONE DELLA TECNICA DI ISPEZIONE PATTERN-BASED PER LA VALUTAZIONE DI USABILITA' DI SISTEMI DI E-LEARNING



Controlled experiment

- Aim
 - Validate the AT inspection
- Evaluation techniques involved
 - Heuristic evaluation
 - Direct observation
 - AT inspection
- The study is illustrated in detail in:

Ardito C., Costabile M., De Angeli A., Lanzilotti R. (2006). Systematic Evaluation of e-Learning Systems: an Experimental Validation. Nordic Conference on Human-Computer Interaction 2006 (NordiCHI 2006), Oslo, Norway, October 14-18, 2006. (pp. 195-202). ACM Press.

Method

- Validation metric was defined along three major dimensions of
 - *Effectiveness*: the number of quality problems discovered by an average inspector, their severity, and consistency across different inspectors
 - *Efficiency*: the time expended in relation to the effectiveness of the evaluation
 - *User satisfaction*: subjective parameters with respect to the evaluation technique and application

Experimental Hypotheses

- Effectiveness Hypothesis
 - AT inspection should increase evaluation effectiveness as compared to heuristic evaluation
- Efficiency Hypothesis
 - no difference in efficiency between the two techniques
- Satisfaction Hypothesis
 - satisfaction of the inspectors that used the AT inspection should be major than the satisfaction of the evaluators that exploited heuristic evaluation

Participants

- 73 senior students of a Human-Computer Interaction class at the University of Bari
- 25 students of another Computer Science class, thus acting as users for the user testing method

Design

- The three evaluation techniques were manipulated between-subjects
- Each group was assigned to one of the three experimental conditions
 - **Heuristic evaluation** group performed the evaluation applying the “learning with software” heuristics [Squires and Preece, 1999]
 - **User testing** group, each evaluator observed a learner who performed a set of tasks with the e-learning application while thinking-aloud
 - **AT inspection** group used the inspection technique with ATs, as proposed by AT inspection

Procedure

- A week before, all participants were presented with a short demonstration of the application to be evaluated, lasting almost one hour
- A couple of days before, a training session of about one hour introduced participants with the conceptual tools to be used during the experiment
- The experiment consisted of two experimental sessions lasting three hours each
 - First session: participants evaluated the e-learning system applying the technique they were assigned to
 - Second session: evaluators were asked to type each detected problem in an electronic form and to fill in the evaluator-satisfaction questionnaire

Data coding...

- Two expert evaluators independently examined all the electronic forms in order to identify single and unique quality problems
 - 247 problems
 - 49 non problems
 - 7% of the problems written by the inspectors applying the heuristic evaluation
 - 3% of the problems written by the inspectors applying the AT inspection or performing user testing

...Data coding

- **Content** included negative statements on the quality of the information provided by the system in terms of clarity and completeness
- **Graphical design** regarded adverse comments on aesthetic aspects of the interface
- **Technical problem** covered issues of pages visualization, compatibility of the system with the browser, and downloading time
- **Feedback** included negative statements addressing the communication between the user and the interface
- **Navigation** covered problems related to the task of moving within the system

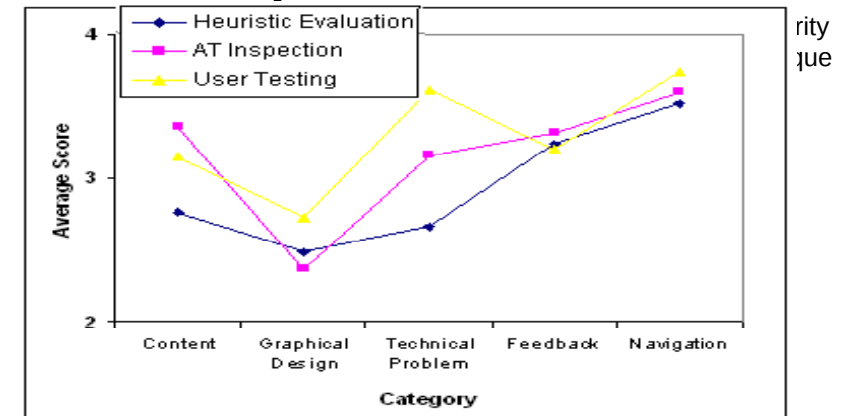
Results...

■ Effectiveness analysis

Category	Heuristic Evaluation		AT Inspection		User Testing	
	Freq.	%	Freq.	%	Freq.	%
Content	34	12	165	31	34	14
Graphical Design	55	20	150	29	33	14
Technical Problem	6	2	25	5	18	7
Feedback	102	36	110	21	66	27
Navigation	84	30	75	14	90	37
Totale	281	100	525	100	241	100

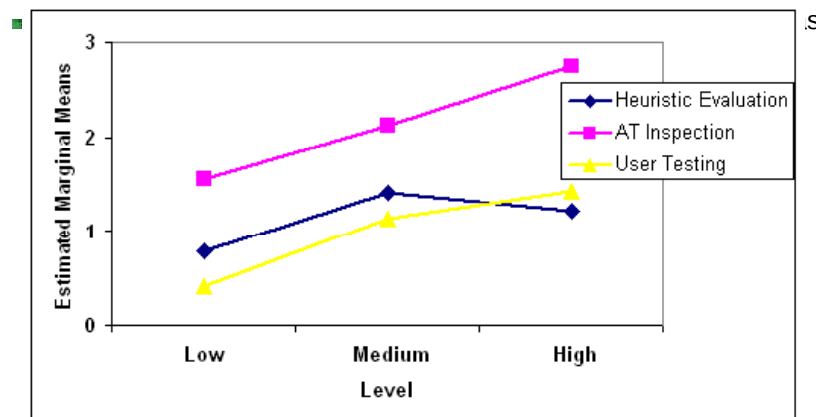
...Results...

■ Effectiveness analysis



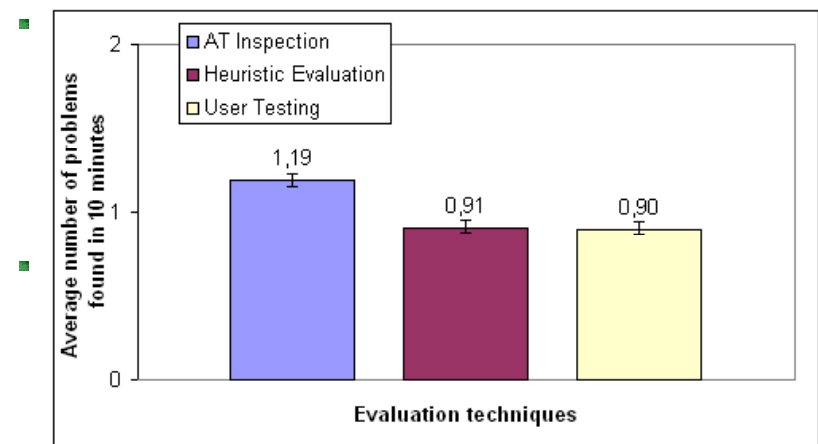
...Results...

■ Reliability defined as consistency of results



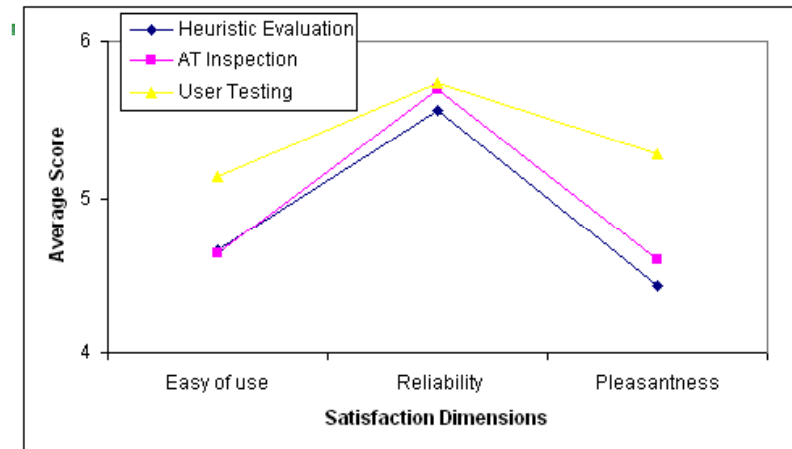
...Results...

■ Efficiency analysis



...Results...

■ Satisfaction analysis



...Results

■ Satisfaction analysis

- Satisfaction with the evaluated application was assessed by a 12-item semantic differential scale
 - An overall index was computed by averaging scores to the individual items

$$F_{(2,69)} = 3.93, p < .05_$$

- Post-hoc analysis applying the Bonferroni's correction, demonstrated that
 - after applying AT inspection participants were more positive in evaluating the application (mean = 4.87)
 - participants in the heuristic evaluation condition (mean = 4.32)
 - participants in the user testing condition expressed an intermediate evaluation (mean = 4.56)

Experimental summary

- The hypotheses of effectiveness and efficiency were supported
 - Participants of the AT inspection group were able to find more problems
 - AT inspection was also more reliable and more efficient
- The study also revealed that AT inspection addresses more specific problems
 - User testing and heuristic evaluation highlighted problems common to all interactive systems
- The hypothesis of satisfaction was not supported
 - The evaluator satisfaction among the three techniques revealed an advantage of user testing as regards to pleasure in using the technique



STUDIO SUL CAMPO PER VALUTARE UN SISTEMA DI M-LEARNING



Explore!

- M-learning system supporting middle school students during a visit to archaeological park
- Implements the excursion-game learning technique on a commercial cell phone
 - Excursion-game is similar to a treasure hunt, students play it in groups
- Improves the visitors' experience at historical sites
 - No site infrastructure required



Evaluation study

- Compare students' experience while playing Gaius' Day in the paper-based version with the experience of the electronic version
- The study is illustrated in detail in:

Costabile M.F., De Angeli A., Lanzilotti R., Ardito C, Buono P, Pederson T. "Explore! Possibilities and challenges of mobile learning". *International Conference on Human Factors in Computing Systems 2008 (CHI 2008)*. Florence, Italy, April 5-10, 2008. ACM Press. (pp. 145-154)
- Several stakeholders involved in the studies
 - students
 - teachers
 - experts of teaching history
 - archaeological park curators
- Pilot study with 24 students to evaluate the system reliability and evaluation methodology

Hypotheses

- Behaviour
 - The sequential nature of the game in Explore! will affect problem-solving strategies
 - students' performance with mobile version could:
 - increase, because players attention is focused on current mission
 - decrease, as fewer cues are available to solve each mission
- Engagement
 - Explore! will be more engaging than the original paper-based game
 - interact with multimedia content during the game and the debriefing
- Learning
 - The Explore! multimedia content will help users to form a clearer mental model of life at ancient Roman times

Participants and design

- Participants
 - 42 students (aged 12) of the School "Michelangelo" in Bari
 - familiar with the use of PCs, cell phones
 - previous experience of the paper-based game played at a different historical park
- Design
 - game type (paper-based vs. mobile) was manipulated between-subjects
 - 19 students, divided into 5 groups, played the paper-based version
 - 23 students, divided into 6 groups, played the mobile version
 - groups were formed by the teacher to guarantee social and cognitive homogeneity

Procedure

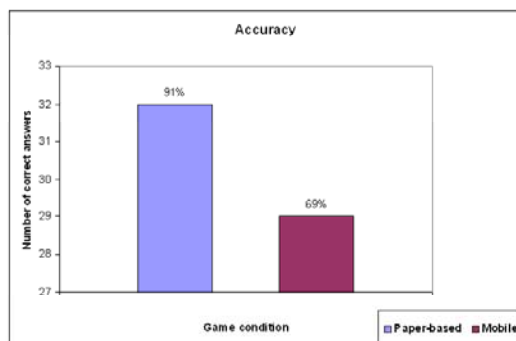
- **Game phase**
 - participants received all the material
 - 3 groups, belonging to the same game condition, played the game simultaneously
 - missions were proposed in different order to avoid interferences among groups
 - 2 experimenters for each group (one in charge of videotaping and one in charge of taking notes)
 - 4 external observers overlooked students at a distance
 - participants answered a questionnaire addressing game experience
- **Debriefing phase**
 - participants were debriefed according their condition
 - they answered a questionnaire regarding learning self-assessment
- **The next day at school**
 - students were administered a test to evaluate the knowledge they had acquired during the game
 - they composed essays and drew pictures about their experience

Instruments

Behaviour	Naturalistic observations Questionnaire Focus group Essays and drawings
Engagement	Open and closed questionnaire Naturalistic observations Focus group
Learning	Observations during the debriefing phase Questionnaire Essays

Behaviour 1/2

- **Accuracy:** mobile condition more prone to errors than the paper-based one, with a difference of some 20 % points
 - the sequential game procedure implemented in the mobile condition
 - the 'Oracle' displayed only the glossary entry directly related to the active mission
- **Speed:** the paper-based condition completed the missions faster



Behaviour 2/2

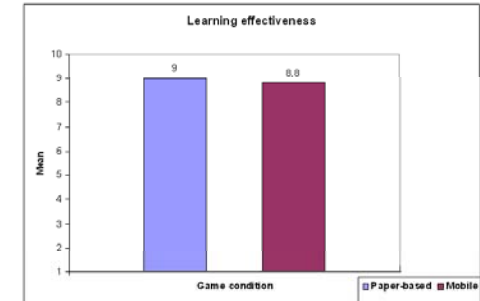
- **Leadership:** leader role emerged in mobile groups
 - no clear trends emerged in the paper-based condition
 - 50% of the participants who played the leader role in the mobile condition happened to be the one holding the cell phone
- **Competition:** mobile groups more competitive than the paper-based ones
 - paper-based groups were more talkative and often engaged in jokes and chit chat when they met
 - mobile groups appeared very concentrated on their tasks and did not want to exchange any comments with their adversaries
 - "Have you found ****?", "Have you finished?", "What mission are you carrying out?"

Engagement

- No differences between the conditions were found (mean = 4.52; mode = 5)
 - three best and the three worst features of the game
 - best features: mobile groups reported more positive features (mean per participant = 2.7) than paper-based groups (mean = 1.9)
 - Paper-based group: place
 - Mobile group: cell phone, 3d reconstruction
 - worst features: no differences emerged (mean = 1)
 - Both groups: complexity

Learning

- **Duration:** the debriefing after the mobile game lasted 10 minutes longer than the debriefing after the paper-based game
 - Time to interact with electronic map and to analyse 3D reconstructions
- **Educational impact:** no significant differences emerged
 - students were very positive and they all agreed they had learned something (mean = 4.1; std dev = 0.66)
- **Educational effectiveness:** no significant differences emerged
 - distributions are strikingly similar



Paper-based game vs mobile game

- **Behavior:** different problem-solving strategies due to the sequential order of game missions
 - *lessons for designers:* mobile games require more interaction freedom and context-dependent information to enhance the overall user experience
- **Engagement:** no difference
 - ↑ students enjoyed the use of cell phone
 - ↑ students appreciated the 3D reconstructions both on phone and during debriefing
- **Learning:** no difference
 - ceiling effect (participants' performance was very high)
 - ↑ students were not distracted by the technology
 - ↑ e-learning is as equally valuable as traditional learning provided that appropriate techniques are used



Grazie per l'attenzione!

Rosa Lanzilotti
 IVU Lab
 Dipartimento di Informatica
 Università di Bari, via Orabona, 4 – 70125 Bari
 lanzilotti@di.uniba.it

