Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea Magistrale in Informatica - Tecniche del Software

APPROCCI RECENTI ALLA VALUTAZIONE DI USER EXPERIENCE E USABILITÀ

Relatore:
Chiar.mo Prof.
Vitali Fabio

Presentata da:
Giardino Di Lollo Chiara

Sessione III
Anno Accademico 2020-2021

Alle difficoltà e agli imprevisti
perché altrimenti che gusto ci sarebbe?

Introduzione

Fin dalla nascita dell'industria si cerca di produrre beni utili e di facile utilizzo per gli utenti finali. Con l'avanzare della tecnologia e l'avvento di internet si è fatto sempre più strada il concetto di esperienza dell'utente e il bisogno di massimizzarlo.
I produttori iniziano così a porsi delle domande: che cosa vuole ottenere l'utente? Qual è il suo background culturale e tecnico? Qual è il contesto d'utilizzo? Che cosa deve fare la macchina e che cosa invece l'utente?
La risposta a queste domande dà forma ad un unico ed importantissimo concetto: l'usabilità, argomento cardine di questa tesi.
La scelta dell'argomento di questa dissertazione è stata fatta in quanto l'usabilità è una disciplina molto affascinante ed importante, sopratutto nell'ambito informatico.
Molto spesso purtroppo però l'importanza dell'usabilità viene accantonata a discapito dei costi che implica.
É molto frequente, infatti, che l'analisi di un sistema sotto l'aspetto dell'usabilità venga evitato adducendo come scusa l'impossibilità di affrontare il dispendio di risorse economiche e di tempo necessarie per un test.
Vengono così rilasciate applicazioni, siti o programmi che, seppur in grado di fruire lo scopo per cui sono state create, per l'utene non rappresentano un sistema semplice e facile da usare e, di conseguenza, lo abbandonano o lo usano svogliatamente.
Seppure effettivamente i test d'usabilità, nella loro versione più elaborata, rappresentano un costo che non tutte le azienda sono in grado si sostene-

re, soprattutto dato che spesso ci si rivolge ad agenzie di consulenza per effettuarle, non tutti sanno che esistono una moltitudine di strumenti molto semplici e gratuiti che possono fornire dei dati chiave per risolvere piccoli o grandi problemi di usabilità di un sistema.
Spesso si da quindi per scontato che eseguire un test d'usabilità non è possibile, senza davvero documentarsi a riguardo.
Questo è stato il punto focale dei miei studi: approfondire al meglio l'argomento ed individuare possibili nuovi approcci.
La dissertazione è il risultato di un ampio lavoro di ricerca ed approfondimento sull'attuale percezione ed utilizzo dell'usabilità sia in campo scientifico sia nel web in generale e, di conseguenza, nella vita di tutti i giorni.
Il punto di partenza è stato quello di documentarsi sulle pubblicazione accademiche più recenti, osservando quali metriche sono utilizzate, attraverso quali strumenti e se vengono effettuate modifiche agli strumenti usati o se invece viene fatto un utilizzo acritico.
Dopo una prima analisi della letteratura degli ultimi 10-15 anni è emerso come la stragrande maggioranza degli articoli analizzati facessero un uso acritico degli strumenti, per il mero scopo di valutare l'usabilità di un sistema e, di conseguenza, non erano espressi commenti, contestazioni o proposte di sviluppi inerenti i questionati o gli strumenti utilizzati in fase di valutazione.
Per questo motivo è stata effettuata una ricerca più radicata: dopo aver individuato alcuni articoli interessanti e rilevanti è stata esplorata tutta la loro bibliografia e, per ogni fonte citata, è stato fatto lo stesso in maniera ricorsiva.
Questa analisi a "cascata" è stata eseguita fino a risalire a quelle che possono essere definite le basi dell'usabilità.
Questo elaborato mira a proporre al lettore una fotografia dell'usabilità attuale, attraverso l'approfondimento di metriche e strumenti utilizzate nel corso degli anni, ad analizzare come negli anni gli strumenti si siano evoluti ed, infine, proporre nuovi spunti su come sfruttare alcuni degli indicatori di usabilità per eseguire test veloci, facili e sopratutto ripetibili nel tempo per

poter raccogliere dati di usabilità in modo continuo e non solo nella fase iniziale di progettazione.
La nascita dell'usabilità si fa risalire agli anni 80-90 ma già prima, addirittura nel I secolo a.C. iniziano ed emergere le prime direttive sull'importanza di rispettare alcune indicazioni fondamentali nell'ideazione e costruzione degli edifici.
Con le guerre, in special modo la prima e seconda guerra mondiale, si ha un'ulteriore spinta alla necessità di produrre armi letali ma soprattutto intuitive nell'utilizzo, che non richiedano un dispendio di energie troppo elevato e sopratutto che siano facilmente utilizzabili anche da persone inesperte.
Si inizia così a parlare di *human factors* e dell'importanza di tenere in considerazione l'utilizzo e l'utente per cui un prodotto è destinato in fase di progettazione e produzione.
Negli anni 80 negli uffici iniziano a diventare sempre più presenti i Personal Computer che, sebbene rappresentano una vera innovazione per il mondo del lavoro, per i lavoratori non diventano altro che fonte di frustrazione ed ansia. I primi PC erano infatti pensati per utenti esperti ed in grado di gestire eventuali problematiche scaturite dall'utilizzo dei computer.
Bisogna tener conto che si trattava di elaboratori distanti anni luce da quelli che utilizziamo oggi, è quindi facile mettersi nei panni di un impiegato dell'epoca e rendersi conto che più che un surplus i computer rappresentavano un ulteriore rallentamento al lavoro di tutti i giorni.
Emergono così i primi studiosi dell'argomento, vengono pubblicati i primi articoli ed inizia ad emergere il termine usabilità.
Proprio in questi anni iniziano ad emergere i grandi nomi di quelli che poi saranno individuati come i padrini dell'usabilità.
Don Norman, Jakob Nielsen e Joe Dumas sono alcuni esempi dei nomi altisonanti del mondo dell'usabilità.
Gli anni 90 e 2000 rappresentano il vero e proprio boom, la spinta decisiva per far ottenere all'usabilità e ai suoi studiosi la notorietà meritata. Questa escalation può essere ricondotta a tre fattori principali; innanzitutto il prezzo

dei PC scende rendendo possibile, a quasi tutte le famiglie, l'acquisto di un computer da tenere in casa; da strumento di lavoro presente solo negli uffici inizia quindi ad arrivare in tutte le case, rispettando la predizione fatta da Bill Gates nel 1980

> "*A computer on every desk and in every home*"

Contestualmente nasce il web, la vera e propria svolta per il mondo dell'usabilità: gli utenti hanno finalmente modo di osservare e valutare diversi prodotti prima di scegliere quale utilizzare o acquistare (negli anni precedenti invece si acquistava un software e solo dopo si scopriva se il suo utilizzo fosse semplice o meno).
Sempre più siti web vengono creati ed emerge l'importanza dell'esperienza di navigazione dell'utente. Infine, ma non meno importante, grazie a una fluttuazione dei titoli NASDAQ in borsa (definita dai tecnici bolla dot-com) il mondo dell'informatica e dell'usabilità ottengono un'esposizione mediatica senza precedenti.
Lo stesso Jakob Nielsen ha dichiarato che, in quel periodo, i media smaniavano per fare intervista con esperti del campo dell'usabilità e lui stesso ha tenuto più di un centinaio di interviste.
Parafrasando le parole di Nielsen

> "*This strong positive PR for UX made many companies think 'we need some of that'*" [59]

la grande esposizione mediatica, unita agli altri fattori, rese l'usabilità diventa uno dei fattori più importanti da tenere in considerazione dal sistema manageriale dei produttori di software e non solo.
Nel capitolo **Storia dell'usabilità** verrà trattato in modo più approfondito questo excursus storico.
Nel capitolo **Usabilità: definizione e metriche** viene formalizzata la definizione di usabilità e vengono analizzate due diverse visioni dell'usabilità.
Viene esaminata quella dell'ISO che definisce l'usabilità basandosi su tre metriche: efficacia, efficienza, soddisfazione.

In contrapposizione Jakob Nielsen basa la propria definizione di usabilità su cinque metriche: apprendibilità, efficienza, memorabilità, errori e soddisfazione.
In questo capitolo viene fatta una classificazione dei diversi test d'usabilità che esistono in base alle loro caratteristiche: test moderati e non modearti, questionari post-task e pst-studio, test delux e gerrilla test.
Viene riproposto e commentato l'iter da seguire per poter somministrare un test d'usabilità, diviso nelle varie fasi: pianificazione, setup, reclutamento, preparazione, esecuzione ed analisi.
Infine vengono analizzate e commentate alcune critiche smosse ai test d'usabilità, spesso usate come scusante per non effettuarne, nel corso degli anni.
Il capitolo **Metodi per misurare l'usabilità** è il frutto della lettura ed analisi di più di 60 articoli accademici il cui scopo era di valutare o confrontare l'usabilità di applicazioni, siti o sistemi in generale.
Analizzando questi articoli sono emersi molteplici tecniche utilizzate dagli studiosi.
Le più citate sono state approfondite, divise per metrica misurata, riproponendo struttura, ideatore, utilizzo ed eventuali criticità e punti di forza.
Le tecniche analizzate che servono a misurare l'efficacia sono il tasso di completamento (*task completion rate*) ed il numero di errori.
Per la rilevazione dell'efficienza sono state approfondite il tempo di completamento dei task (*task time*) ed il numero di click.
Il carico di lavoro mentale, invece, viene analizzato attraverso l'utilizzo di questionari quali l'*Overall Workload* (OW), il *NASA Task Load Index* (NASA-TLX) ed il *Subjective Workload Assessment Technique* (SWAT).
Per quanto riguarda la soddisfazione sono stati approfonditi l'*After-Scenario Questionnaire* (ASQ), il *Computer System Usability Questionnaire* (CSUQ), la *System Usability Scale* (SUS), l'*Usability Metric For User Experience* (UMUX) e la sua versione ridotta (UMUX-LITE).
In questo capitolo è interessante notare come, quasi sempre, con il passare degli anni si cerca di analizzare gli studi pregressi ed elaborare un nuovo

questionario costituito da meno quesiti ma che comunque fornisca risultati soldi e affidabili quanto questionari ben più corposi.
Lo si può notare in modo lampante confrontando, per quanto riguarda l'analisi della soddisfazione, il CSUQ (1995, 19 quesiti), la SUS (1996, 10 quesiti), l'UMUX (2010, 4 quesiti) e l'UMUX-LITE (2013, 2 quesiti).
Un altro dato che emerge dall'osservazione dei vari questionari è che, molto frequentemente, vengono sfruttate delle scale Likert (scale bipolari rappresentato da un numero arbitrario di risposte accompagnate da etichette descrittive) come risposte a quesiti del questionario; questo rende le risposte confrontabili e, di conseguenza, i questionari più facilmente riassumibili in un punteggio finale, solitamente risultante da equazioni che mettono in relazione domande e risposte fornite.
Ovviamente esistono una moltitudine di questionari ed indicatori utilizzabili per valutare l'usabilità altrettanto validi tuttavia, per evitare di trasformare questa dissertazione in un manuale delle tecniche per misurare l'usabilità, sono stati scelti quelli più citati e che comunque rappresentano un'evoluzione di strumenti equivalenti già esistenti.
Una volta analizzati, per metterli a confronto sotto l'aspetto delle citazioni (e di conseguenza per catturare la mole d'utilizzo), sono stati raccolti dati inerenti il numero di risorse presenti in rete su due dei motori di ricerca più conosciuti: Google e Google Scholar.
La scelta è stata fatta perché su Google Scholar sono presenti milioni di risorse nell'ambito della letteratura accademica; in contrapposizione Google rappresenta il motore di ricerca più utilizzato e che consente, tramite la ricerca per parole chiave, di avere accesso a documenti, siti e qualsiasi risorsa presente in rete.
Utilizzando Google Scholar e Google lo scopo è quello di catturare il numero di risorse presenti in ambito accademico ed in ambito più generale.
Va infatti tenuto conto che su internet sono presenti moltissimi siti in cui viene spiegato in breve cos'è l'usabilità e come eseguire dei test; per questo motivo, nonostante l'importanza degli articoli accademici, è molto più pro-

babile che un programmatore, o chicchessia, inesperto in materia e con poco tempo a disposizione per imparare delle nozioni basilari si rivolga a Google per cercare in modo veloce ed efficace come testare il design del prodotto che si sta sviluppando, piuttosto che dedicare molto tempo ad una ricerca approfondita su Google Scholar o siti equivalenti.
Oltre ad una suddivisione della ricerca tra i due motori di ricerca è stata fatta una frammentazione temporale.
Per entrambi i motori di ricerca è stata fatta sia una ricerca generale (senza vincoli), sia una ricerca in cui sono filtrati i risultati in modo tale da ottenere le risorse presenti in rete solo dal 2010 ad oggi.
Questa scelta è stata fatta partendo dal presupposto che alcuni dei questionari risalgono ai primi anni 80 mentre i più recenti risalgono anche al 2010.
In questo modo i risultati delle ricerche mostrano il numero di citazioni generali (indipendentemente dalla data di pubblicazione) ed il numero di citazioni a partire da una data "neutra" (in cui quindi tutti i questionari sono già stati pubblicati), una sorta di punto di partenza semi neutrale per tutti gli indici.
Ovviamente la ricerca per parole chiave, sopratutto su google, può produrre risultati forvianti; basta pensare che SUS vuol dire "suo" in spagnolo, è il cognome di molte persone, è l'acronimo del servizio unico dei servizi ecc; allo stesso modo SWAT è l'acronimo del corpo speciale americano.
Per questo motivo spesso in fase di ricerca è stato aggiunto il suffisso "usability", per cercare di escludere tutti i risultati non inerenti; ovviamente questo escamotage non elimina del tutto i risultati non coerenti con la ricerca.
Da una prima analisi è emerso come i questionari più citati siano il NASA-TLX (misuratore di carico di lavoro mentale) e la SUS (indicatore di soddisfazione percepita).
Dopo aver analizzato gli indicatori singolarmente sono state effettuate delle ricerche, sempre per parole chiave, in cui sono stati messi in relazioni a coppie i vari questionari.
Anche in questo caso la ricerca è stata fatta parallelamente su Google e Google Scholar ma senza fare la distinzione riguardo la data di pubblicazione

delle risorse.
Come prevedibile la coppia che ha avuto come risultato il numero maggiore di citazioni è stata quella SUS NASA-TLX.
É infatti emerso come spesso per valutare l'usabilità dei sistemi si analizzi la soddisfazione ed il carico di lavoro mentale percepito dagli utenti.
I risultati più dettagliati si trovano nel capitolo **Statistiche di utilizzo degli indicatori di usabilità**.
L'ultimo capitolo della tesi, **Nuovi approcci**, rappresenta un'analisi di come sono percepiti oggi i test d'usabilità e quali sono i nuovi strumenti per misurarla.
Nell'epoca del web 3.0 del 5G e degli smartphone l'usabilità si scontra con la necessità di produrre tanto e velocemente.
Oggi, ogni giorno, vengono pubblicate una moltitudine di applicazioni, siti, programmi ecc.
In un certo senso si può dire che attualmente l'importante è avere una buona idea e renderla pubblica il prima possibile, prima che qualcun'altro si impossessi della stessa intuizione.
In un mondo in cui il tempismo è tutto, trovare del tempo in cui effettuare anche un semplice test d'usabilità per qualcuno può rappresentare, purtroppo, una perdita di tempo.
Nascono così o, in alcuni casi, vengono rispolverati metodi per analizzare l'usabilità efficienti ma con una peculiarità: sono composti da un solo quesito.
É questo il caso del *Subjective Mental Effort Question* (SMEQ), dell'*Usability Magnitude Estimation* (UME) e del *Single Ease Question* (SEQ).
Questi tre indicatori sono poco utilizzati per quanto riguarda la ricerca accademica tuttavia, cercano sul web combinazioni come "usability test" "how to measure usability" molti articoli su siti dedicati all'usabilità (o comunque al mondo informatico) li indicano come alcuni degli strumenti più efficienti e veloci per misurare l'usabilità.
Cercando articoli accademici a riguardo, dato che l'affidabilità dei siti web è sempre effimera, è emerso come effettivamente l'UME, il SEQ ed il SMEQ

siano effettivamente una valida alternativa ad indicatori più noti e sostanziosi quali SUS, UMUX, ecc.
Nasce spontanea una domanda: se basta davvero una singola domanda per catturare l'usabilità percepita perché non viene posta? E sopratutto perchè non porla più volte durante il ciclo di vita del sistema?
Molte applicazioni durante l'utilizzo chiedono feedback all'utente; ne è un esempio UberEats (app di food delivery) che dopo ogni ordine chiede di valutare il rider, il ristorante ed i singoli piatti ordinati; analogamente YouTube, di quando in quando, durante le pubblicità all'interno dei video chiede all'utente di scegliere tra delle macro categorie di argomenti a cui è interessato.
Se già durante l'utilizzo di diversi siti web ed applicazioni vengono proposte all'utente domande per ottenere informazioni sul marketing o per dare valutazioni agli iscritti alla piattaforma allora perché non fare lo stesso per valutare l'usabilità?
Questa potrebbe essere la risposta alla mancanza di test d'usabilità dovuta alla scarsità di risorse.
In questo modo, infatti, se non si dispone di risorse per effettuare i test in fase di realizzazione del progetto, si potrebbe chiedere all'utente dopo il primo utilizzo come valuta l'esperienza di utilizzo.
Lo si potrebbe fare sia in generale, sia in termini di singole azioni compiute, o task (l'acquisto di un prodotto, l'invio di una mail, l'aggiunta di un elemento in wishlist, ecc).
Questo approccio, inoltre, permetterebbe di misurare l'usabilità in modo ripetuto nel tempo: misurare se l'esperienza dell'utente migliora man mano che usa il sistema ed inoltre valutare se eventuali aggiornamenti aumentano o diminuiscono la percezione di usabilità.
Cercando accenni a questo tipo di misurazione, in articoli accademici o in generale nel web, si ottiene poco e nulla.
Questo non implica in modo inoppugnabile che non vengono effettuati test d'usabilità nel corso del tempo per una stessa applicazione ma semplicemente che non vengono pubblicati articoli a riguardo; infatti può essere una pratica

effettuata ma i cui dati e risultati non vengono condivisi dalle aziende anche per questioni concorrenziali.

Sarebbe tuttavia interessante analizzare se e come la percezione di soddisfazione, il carico di lavoro mentale e tutti gli aspetti dell'usabilità dei sistemi varia nel tempo, non solo fermandosi all'analisi iniziale del sistema in quanto dato interessante in fase di presentazione e lancio sul mercato ma come dato da usare come spunto per migliorie e nuove implementazioni.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Storia dell'usabilità

1.1 I primi approcci al design ed alla user interface

1.1.1 Le origini

I primi approcci al design vengono ricondotti al I secolo a.C. quando Marco Vitruvio Pollione, architetto e scrittore romano, nella sua opera "De architectura"[87] (opera composta da 10 libri sull'architettura) definì i requisiti che ogni costruzione avrebbe dovuto rispettare:

- **utilitas**: utilità nella funzione a l'edificio è destinato;
- **firmitas**: solidità nella statica e nei materiali;
- **venustas**: bellezza, estetica e venustà (bellezza in cui l'armonia delle proporzioni, accompagnata da un'ineffabile grazia, ispiri il senso di una ideale perfezione).

L'opera di Vitruvio fu di grande ispirazione: molti, tra cui Leonardo Da Vinci, presero spunto dalle sue indicazioni per realizzare opere. Ne è un esempio l'uomo Vitruviano di Da Vinci che è stato realizzato misurando e calcolando le proporzioni del corpo umano.

Vitruvio, in un certo senso, può essere considerato uno dei primi studiosi di ergonomia ed usabilità.

1.1.2 Human factors ed ergonomia

Uno degli ambiti a cui si riconduce la nascita dell'usabilità è sicuramente quello militare e dell'*human factors*; in ambito militare uno degli scopi principali è mettere gli uomini in condizione di riuscire a utilizzare al meglio ed in modo efficiente gli strumenti a disposizione.
Durante la prima e la seconda guerra mondiale molte risorse furono impiegate per rendere il design delle armi il più funzionale possibile. La necessità all'epoca era di avere armi precise, semplici da usare ed efficaci nella neutralizzazione del nemico.
É proprio in questo ambito che emergono quelle che possono essere definite le prime metriche dell'usabilità, sebbene mai formalizzati e specifici per il campo militare[80]:

- **apprendibilità**: quanto è veloce il processo di apprendimento? (il ricambio dei soldati era ingente e il tempo per formare i nuovi cadetti poco);
- **esperienza richiesta**: quanto influisce l'esperienza nel rapporto tempo/esplosione di colpi?
- **efficienza ed efficacia**: quali cambiamenti al design delle armi le renderanno più letali?
- **sforzo**: come è possibile modificare il design per rendere lo sforzo necessario all'utilizzo minimo?

Inizia così ad emergere la consapevolezza che analizzando ed intervenendo sul design degli strumenti si possono ottenere risultati migliori.
Il termine ergonomia deriva dal greco *érgon* (lavoro) e *nomos* (regola, legge), utilizzato per la prima volta nel 1857 da Wojciech Jastrz?bowski in un

giornale polacco. Negli anni 50 Murrell riprese il termine e lo utilizzò per definire le linee guida nel design di prodotti, servizi o ambienti per andare incontro alle necessità degli utenti[89].
Si inizia così ad avvertire la necessità di analizzare le necessità degli utenti, di studiare come migliorare l'esperienza degli utenti; nascono così termini come *user experience*. Si inizia così a gettare le basi per una nuova disciplina: l'usabilità.

1.2 1980: l'avvento dei PC, HCI ed usabilità

Con il diminuire dei prezzi dei computer negli anni 80 inizia la vera e propria ascesa dell'usabilità. In quegli anni gli impiegati iniziano ad avere i loro PC nonostante la loro conoscenza a riguardo fosse molto limitata se non, quasi sempre, nulla.
I programmatori, all'epoca, creavano software dando per scontato che gli utenti fossero formati, che comprendessero l'architettura dei software e che avessero dimestichezza con i computer in generale.
Questa ipotesi errata portò velocemente ad un punto critico: i PC venivano visti dagli utenti come fonte di frustrazione ed ansia data la difficoltà di interazione.
É in questo momento che nasce la necessità di studiare l'interazione tra uomo e computer (*human-computer interaction* o HCI) ed anche l'usabilità; per i designer ed i programmatori divenne uno degli obbiettivi principale quello di realizzare prodotti non pensati per utenti specialisti ma per l'utente generico.
Nel corso degli anni 80 molti studiosi si occupano dell'argomento; più nel dettaglio possiamo analizzare alcuni eventi significativi[78]:

- nel 1982 Clayton Lewis pubblica un report tecnico dell'IBM in cui viene utilizzato il "Thinking Aloud"(ideato da Herb Simon) nel design delle interfacce;

- nel 1983 viene pubblicato "The Psychology of Human Computer Interaction" in cui vengono descritti alcuni trumenti per analizzare l'usabilità (GOMS e Keystroke Level Modeling);
- nel 1984, durante il Super Bowl, l'Apple introduce il Macintosh utilizzando la semplicità di utilizzo come fulcro della pubblicità;
- nel 1984 Harry Hersh e Dick Rubinstein pubblicano per la prima volta un intero libro dedicato alla HCI (The Human Factor);
- nel 1985 viene pubblicato "Designing for Usability: Key Principles and What Designers Think." in cui viene discussa l'importanza di una continua attenzione dedicata agli utenti, di misurazione empirica e progettazione iterativa;
- nel 1986 John Brooke idea la SUS, uno dei questionari attualmente più utilizzato;
- nel 1987 viene pubblicato "Designing the User Interface", un libro in cui vengono espressi i principi pratici e le linee guida per sviluppare un'interfaccia di alta qualità, percepita dagli utenti come comprensibile e controllabile.

1.3 1990: la nascita dell'usabilità moderna

Gli anni a cavallo tra la fine del 1980 e l'inizio degli anni 90 possono essere definiti gli anni in cui l'usabilità diventa una vera e propria disciplina e professione come ha dichiarato Joe Dumas, uno dei fautore e contributore dell'usabilità[78].
Sono questi gli anni cruciali: vengono pubblicati i primi ISO standard in cui si parla di usabilità e delle metriche che la compongono, vengono pubblicati moltissimi studi in cui si definiscono strumenti per analizzare l'usabilità (SUMI, TAM, PSSUQ, ecc) evengono pubblicati studi e opere a riguardo. Inizia quindi il periodo d'oro dell'usabilità, con un incremento vertiginoso

delle persone che dedicano i loro studi e le loro carriere a questa disciplina.
Non è un caso che il boom sia avvenuto negli anni 90: nello stesso periodo nasce e si diffonde il web, nelle case inizia ad essere presente un PC ed una connessione alla rete; i software ed i siti web, inizialmente utilizzati da una nicchia di persone solo a livello lavorativo, diventano uno strumento utilizzato da tutti ed in costante crescita.
Emerge in questo periodo l'importanza dell'usabilità: fino a prima del web gli utenti acquistavano un software e solo successivamente lo installavano ed utilizzavano. Se il sistema fosse quindi usabile o meno lo si scopriva solo in un secondo momento, in una dinamica *payment first, user-experience second*. Con il web la situazione si ribalta: l'utente apre una pagina web e solo se è comprensibile e facile da utilizzare vi rimane e la utilizza, altrimenti cerca un altro sito equivalente ma più user friendly. In questo caso la dinamica è *user-experience first, payment second*. Per le imprese diventa quindi prioritario rendere i propri sistemi usabili ed appetibili.
Un altro fattore che diede una grande spinta all'usabilità fu la *bolla delle Dot-com* (come conseguenza di una bolla speculativa l'indice NASDAQ arrivò ad un picco tra il 1997 e il 2000) che diede un'enorme visibilità mediatica all'usabilità rendendola l'attrattiva del momento[59].

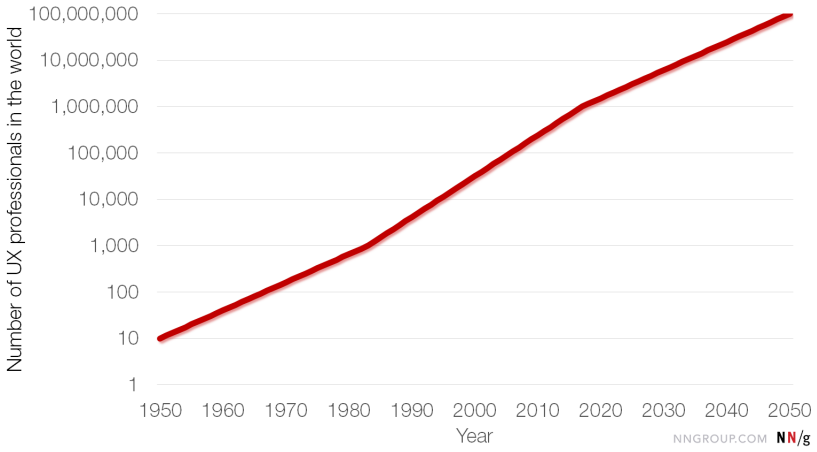


Figura 1.1: aumento dei professionisti che si dedicano all'usabilità nel corso degli anni e predizione sul futuro

Possiamo quindi concludere che la combinazione della nascita del web, dell'esposizione mediatica e della diffusione dei PC hanno dato un impulso molto importante all'affermarsi dell'usabilità come punto cardine dello sviluppo di sistemi e prodotti.

Capitolo 2

Usabilità: definizione e metriche

2.1 Definizione di usabilità

L'usabilità può essere definita come lo strumento attraverso il quale è possibile analizzare il grado di interazione tra l'uomo ed un sistema; rappresenta uno dei parametri fondamentali per l'analisi dell'user experience e, di conseguenza, uno degli aspetti più importanti da tenere in considerazione durante le fasi di progettazione, realizzazione e test di un prodotto, sia esso software o hardware.
Molteplici sono le definizioni di usabilità, simili nell'abito di applicazione e definizione generale, differiscono solitamente per gli aspetti che, secondo l'una o l'altra fonte, contribuiscono all'analisi dell'usabilità stessa. Le definizioni più conosciute sono sicuramente quelle dell'ISO e di Nielsen.

2.1.1 La definizione di usabilità dell'ISO

L'ISO/IEC 9126 *"Information technology - Software product evaluation - Quality characteristics and guidelines for their use"*[39] definisce l'usabilità come

> "la capacità del software di essere compreso, appreso, usato e gradito dall'utente quando usato in determinate condizioni"

L'ISO 9241-11 *"Ergonomic requirements for office work with visual display terminals - Guidance on usability"* affina ulteriormente la definizione di usabilità

> "il grado in cui un prodotto può essere usato da specifici utenti per raggiungere specifici obiettivi con efficacia, efficienza e soddisfazione in uno specifico contesto d'uso"[38]

L'ISO 9241-11 definisce inoltre gli aspetti secondo cui analizzare l'usabilità:

- **efficacia**: capacità degli utenti di completare compiti e di raggiungere gli obbiettivi preposti;
- **efficienza**: sforzo richiesto dagli utenti per completare compiti a loro assegnati;
- **soddisfazione**: facilità d'uso percepita dagli utenti.

2.1.2 La definizione di usabilità di Nielsen

Lo studioso e scrittore Jakob Nielsen nel 1993 scrisse

> "usability in the design of a system lies within the concept of utility and complementary to usefulness"[40]

sottolineando che l'utilità di un sistema è rappresentata dalla sua correttezza funzionale (*usefulness*) e dalla correttezza non-funzionale (*usability*). Nielsen enuncia che l'usabilità è misurabile sotto cinque diversi aspetti, e non tre come definito dall'ISO.
Più precisamente, secondo Nielsen[40], l'usabilità è definita dalle seguenti componenti:

- **apprendibilità**: facilità con cui gli utenti, al primo utilizzo, riescono a svolgere un'attività di base;

- **efficienza**: facilità di utilizzo da parte dell'utente una volta appreso il design del sistema;
- **memorabilità**: capacità degli utenti di utilizzare nuovamente il sistema dopo un periodi di inutilizzo;
- **errori**: errori commessi dagli utenti in fase di utilizzo, tenendo in considerazione la gravità e di facilità di recupero;
- **soddisfazione**: quanto il sistema risulta piacevole agli utenti.

2.2 I test di usabilità

2.2.1 Tipologie di test

É possibile classificare i test d'usabilità in base ad alcune caratteristiche che li contraddistinguono. In base alla modalità di svolgimento del test possiamo distinguere[55]:

- **test moderati**: il test è svolto in presenza di un moderatore che si può interfacciare con i partecipanti (può descrivere lo scenario ed i task, osservare i partecipanti durante l'esecuzione, approfondire comportamenti inattesi); questa modalità è più costosa in termini di tempo tuttavia permette di approfondire alcuni aspetti del sistema in utilizzo fornendo non solo un risultato del test ma anche feedback e informazioni aggiuntive non rilevabili altrimenti;

- **test non moderati**: i partecipanti svolgono il test in modo autonomo seguendo le indicazioni riguardo a task e modalità di svolgimento ricevute dagli organizzatori; si tratta di una modalità molto rapida e scalabile in quanto gli operatori devono occuparsi solamente di inviare il materiale necessario per mettere i partecipanti in condizione di svolgere il test e poi analizzare i risultati ottenuti;

In basa al momento in cui il test viene somministrato ai partecipanti del test è possibile distinguere tra[35]:

- **questionari post-task**: si tratta di test somministrati per la valutazione della percezione degli utenti riguardo singole attività; se lo studio è composto da più task verrà somministrato un test per ogni task previsto;

- **questionari post-studio**: si tratta di test volti alla valutazione complessiva che gli utenti hanno rispetto ad un sistema (prodotto, sito web, app, ecc); vengono somministrati alla fine di una sessione di utilizzo del sistema o dopo che l'utente ha avuto modo di interagire con un prodotto.

In base alle risorse impiegate, in termini economici e di tempo, possiamo distinguere tra[64]:

- **test d'usabilità tradizionali** (o *deluxe usability testing*), si tratta di test ben formalizzati e strutturati, somministrati a circa 100 partecipanti (remunerati) per ottenere dei risultati statistici solidi e la verifica del rispetto dei requisiti del sistema;

- **guerrilla test** (o *discount usability testing*), si tratta di uno studio informale somministrato ad un campione molto ridotto di partecipanti (3-4); questa tipologia di test viene svolta in fase di produzione per identificare i problemi del sistema e risolverli contestualmente; i risultati di questa tipologia di test vanno considerati come indicativi.

2.2.2 Realizzazione di un test d'usabilità

Per poter eseguire un test d'usabilità è necessario seguire delle fasi, indipendentemente dalla tipologia di test scelto è possibile individuare 6 step principali[70]:

1. **pianificazione**, in questa fase bisogna:

   - definire l'assunzione di partenza e gli obbiettivi del test;
   - stabilire le modalità di esecuzione del test;

- identificare le metriche che si vogliono analizzare e gli strumenti attraverso cui farlo;

2. **setup del test**, questa fase consiste:
   - nell'identificazione dei task che costituiranno il test;
   - nella stesura dello script (una guida da seguire durante il test dove sono indicati i quesiti e le domanda da fare ai partecipanti e altri dettagli importanti del test);
   - nel test dello script;
3. **reclutamento dei partecipanti**, ovvero:
   - stabilire una strategia di reclutamento (identificare il target degli utenti che si desidera avere nel test e le modalità di reclutamento);
   - eseguire il reclutamento, fino al raggiungimento del numero di partecipanti prefissato;
4. **preparazione dell'ambiente del test**: che si tratti di un test moderato o non moderato è necessario preparare la stanza o l'ambiente virtuale in cui il test verrà somministrato ai partecipanti;
5. **esecuzione del test**
6. **analisi e revisione**: una volta terminata la sessione di test si può passare all'analisi dettagliata dei risultati del test (e delle eventuali questioni sollevate dai partecipanti nel caso di un test moderato).

2.2.3 Critiche

I test di usabilità, nel corso degli anni, hanno sollevato diverse critiche da parte di utilizzatori e studiosi. Spesso figlie di una conoscenza superficiale dell'usabilità e delle tecniche per valutarla, le critiche riguardano[2]:

- **il costo dei test**: spesso viene sollevata la questione dell'onerosità dei test d'usabilità, tuttavia questo è vero solo se ci si rivolge ad aziende specializzate; spesso si ignora, o si trascura, che è possibile ottenere buoni risultati con test molto economici (basta pensare a questionari online, sondaggi inviati via mail, ecc);

- **la difficoltà di raccogliere dati affidabili con un campione ristretto**: per condurre un test i cui risultati siano affidabili è necessario un campione medio grande; ciò non toglie che sia possibile eseguire test con un campione ristretto di partecipanti ottenendo ugualmente buoni dati ed informazioni inerenti all'usabilità del sistema in analisi;

- **il tempo richiesto**: l'iter da seguire per la creazione e la somministrazione di un test, sebbene non particolarmente elaborato, richiede tempo sia in fase di analisi sia in fase di esecuzione; tuttavia alcuni degli strumenti più comuni per misurare l'usabilità sono costituiti da un set di poche domande che si possono somministrare durante la filiera produttiva alle persone che testano il sistema nella sua funzionalità generale; l'idea comune è che per condurre un test d'usabilità sono necessari giorni o settimane quando a volte basta qualche ora;

- **la poca utilità per i nuovi prodotti**: spesso si pensa che sia inutile eseguire test d'usabilità su nuovi prodotti o sistemi in quanto non ci sarebbe modo di paragonarli con prodotti simili; tuttavia lo scopo primario di un test d'usabilità è quello di valutare, secondo le metriche prescelte, il sistema e solo successivamente effettuare confronti con prodotti simili;

- **l'inutilità nell'individuare la causa**: spesso si sottolinea come i test d'usabilità catturino dei problemi senza però fornire indicazioni sulla loro origine e sulle possibili soluzioni; anche in questo caso si tratta di una critica superficiale in quanto molte tipologie di test prevedono sezioni dove i partecipanti possono descrivere i problemi riscontrati

e quindi analizzando i feedback è possibile intervenire per rendere il sistema più usabile;

- **la poca comprensibilità**: spesso si evitano i teest d'usabilità indicando come motivazione la poca comprensibilità ed il poco interesse da parte della sezione manageriale dell'azienda; questo dato è piuttosto circostanziale in quanto se proposti in un maniera semplice e chiara i risultati dei test sono facilemnte comprensibili ed apprezzabili da chiunque.

Capitolo 3

Metodi per misurare l'usabilità

Negli ultimi 40 anni sono stati svolti molti studi inerenti l'usabilità e, in particolare, studi volti a formalizzare dei modelli per valutare l'usabilità di un sistema sotto le diverse metriche descritte nelle sezioni precedenti.
Molteplici sono i modelli individuati dagli studiosi, i primi teorizzati già dagli anni 70-80 e successivamente utilizzati, confrontati e confermati.
Nella letteratura attuale è piuttosto evidente come, alcuni metodi, siano preferiti ad altri vista la solidità e facilità di utilizzo.
In questo capitolo verranno analizzate le tecniche più note ed utilizzate per analizzare l'usabilità, nel capitolo **Nuovi approcci** alcuni di questi strumenti verranno confrontati con tecniche più recenti per cercare di trarre spunti ed idee su come sfruttare al meglio la molteplicità di strumenti a disposizione.

3.1 Metodi per valutare l'efficacia

3.1.1 Tasso di completamento

Il tasso di completamento (o *completion rate*) rappresenta uno degli indicatori più utilizzati dell'usabilità. L'importanza di questo indicatore è rappresentata dal fatto che non è importante analizzare soddisfazione ed usabilità se l'utente non riesce a completare i task.
É un indice molto semplice da comprendere ed utilizzare: si assegna un pun-

teggio paria 1 se l'utente riesce a completare il task assegnato, 0 in caso di fallimento[57].

$$efficacia = \frac{\text{numero di task completati con successo}}{\text{task sottoposti ai partecipanti}} \times 100\%$$

Sebbene il tasso di completamento a cui puntare è il 100%, secondo Jeff Sauro[75], un tasso superiore al 78% rappresenta la soglia da superare per poter dire che il sistema in analisi è efficace.

3.1.2 Numero di errori

L'analisi del numero di errori commessi dagli utenti in fase di test è un sistema diagnostico che, sebbene richieda più tempo rispetto ad altre tecniche, risulta molto utile per individuare lacune e punti critici del sistema.
Con il termine *errori* si intende azioni involontarie, errori o omissioni non intenzionali che un utente compie durante l'esecuzione di un task[57].
Questa tecnica consiste nel contare il numero di errori commessi, in fase di test, nel tentativo di completare il task; ogni errore deve essere corredato da una breve descrizione, una valutazione di gravità e l'indicazione di appartenenza ad una categoria. Le categorie a cui, generalmente, vengono ricondotti gli errori sono le seguenti[77]:

- **slips**: piccoli errori involontari causati spesso da disattenzione (ad esempio digitare in modo incorretto un dato da inserire); quando molti errori di questa tipologia vengono commessi è opportuno intervenire su campi obbligatori (se si tratta di siti web) e sulla riduzione al minimo indispensabile dei dati da inserire;

- **mistakes**: si tratta in questo caso di dati incorretti inseriti, spesso imputabili alla mancanza di controlli, auto-formattazioni o indicazioni che impediscano all'utente di cadere in questi errori (ne è un esempio l'inserimento della data di nascita nel formato sbagliato in form online);

- **problemi d'interfaccia**: si verificano quando un utente non comprende l'interfaccia ed il design del sistema (ad esempio si fa clic per tornare

indietro in una sezione non cliccabile); in questo caso si può intervenire rendendo l'interfaccia più intuitiva;

- **errori di scenario**: si tratta di errori commessi dall'utente dovuti a escamotage utilizzati in fase di test per simulare alcune azioni (come la simulazione di un pagamento online in fase di test di un e-commerce); l'utente si trova in una situazione a lui sconosciuta o anomala e quindi facilmente commetterà errori.

Generalmente, dato che è possibile commettere più errori durante l'esecuzione del task, si tiene in considerazione semplicemente se l'utente ha commesso errori (dando come punteggio 1) o se non ne ha commessi (punteggio 0) durante i test; questa semplificazione viene fatta per poter ottenere un punteggio finale coerente e non falsato: per analizzare gli errori nel dettaglio, e quindi individuare i punti critici, basterà consultare la documentazione dettagliata del test.

$$\%\text{ di errore} = \frac{\text{numero di errori commessi nell'esecizione del task}}{\text{numero di tentativi di esecuzione del task}}$$

É stato dimostrato che il tasso di errori è correlato al tempo di completamento di un task, al tasso di completamento ed alla soddisfazione rendendo gli errori la motivazione che spiegano valori non positivi degli indicatori appena citati[77].
Secondo un articolo di Jeff Sauro, basato su uno studio che analizza 719 task, il numero medio di errori per task è di 0.7 (tenendo conto che 2 utenti su 3 commettono errori). Solo il 10% dei task osservati è stato concluso senza commettere errori, concludendo che è perfettamente normale per gli utenti commettere errori[76].

3.2 Metodi per valutare l'efficienza

L'efficienza mette in relazione il livello di efficacia raggiunto con il dispendio di risorse. Può essere misurata in termini di tempo, costo economico o

sforzo fisico e mentale.
Per l'utente il tempo che impiega per completare un task e lo sforzo richiesto rappresentano le risorse consumate, in base ad esse si possono distinguere due tipi di efficienza:

- **efficienza temporale**, il cui focus è il tempo necessario per il completamento di un task;
- **efficienza umana** o carico di lavoro mentale (questa caratteristica verrà approfondita nel dettaglio nella sezione **Metodi per valutare il carico di lavoro mentale**)

Per quanto riguarda l'efficienza percepita dall'organizzazione che sottopone task agli utenti si parla di **efficienza economica**, che ingloba il costo dell'utente (ovvero la retribuzione percepita per eseguire il task), il costo delle risorse e delle attrezzature utilizzate e il costo della formazione fornita all'utente[9].

3.2.1 Tempo di completamento dei task

Uno degli indicatori più utilizzati per misurare l'efficienza temporale è il calcolo del tempo necessario per completare con successo un task (*task time*) in termini di minuti e/o secondi[57].
Ci sono due tecniche per analizzare l'efficienza sotto questo aspetto[3]:

- **time-based efficiency**, ovvero un indicatore di quanti task vengono completati con successo in relazione alla misura di tempo scelta (minuti o secondi);

$$\textit{time-based efficiency} = \frac{\sum_{j=1}^{R}\sum_{i=1}^{N}\frac{n_{i,j}}{t_{i,j}}}{N \times R}$$

  dove:

  - N = numero di task da completare;
  - R = numero di partecipanti al test;

  - $n_{i,j}$ = la capacità dell'utente j di superare il task i (vale 1 se il task è stato superato con successo 0 altrimenti);
  - $t_{i,j}$ = il tempo impiegato dell'utente j per superare il task i.

- **overall relative efficiency**, ovvero il rapporto tra il tempo impiegato dagli utenti che hanno completato con successo il task rispetto al tempo totale impiegato da tutti gli utenti.

$$\textit{overall relative efficiency} = \frac{\sum_{j=1}^{R}\sum_{i=1}^{N} n_{i,j}t_{i,j}}{\sum_{j=1}^{R}\sum_{i=1}^{N} t_{i,j}} \times 100\%$$

  con R, N, $n_{i,j}$, $t_{i,j}$ definiti come sopra.

3.2.2 Numero di click

Per quanto riguarda i siti web ed i software esiste un ulteriore tecnica per analizzare l'efficienza: l'analisi del numero di click necessari per eseguire un task.
Nel corso degli anni, infatti, si è cercato di ridurre al minimo il numero di click necessari nell'utilizzo di software e siti web sia per rendere più semplice il task stesso, sia per ridurre il tempo necessario per completarlo.
Spesso si utilizza il conteggio dei click piuttosto che analizzare il *task time* per una questione di semplicità e velocità: basta contare i click necessari senza dover prendere i tempi e calcolare formule matematiche, il tutto senza il bisogno di utenti (il programmatore stesso può effettuare questo conteggio)[74].
Nell'ottica di ottimizzare l'efficienza occorre tuttavia tenere in conto che rendere minimo il numero di click non sempre implica rendere il sistema ottimo in termini di usabilità; spesso infatti per rendere un task eseguibile in un solo click si agglomerano molte informazioni in un'unica schermata ottenendo l'effetto contrario.
Sauro nel suo articolo ha analizzato i dati relativi a 1228 utenti che eseguivano 19 task (per un totale di 4892 osservazioni) su cui ha condotto un'analisi per verificare la correlazione tra *task time* e numero di click.

Dall'analisi è risultata una correlazione di 0.5 e da ulteriori approfondimenti è emerso che il numero di click spiega il 25% del tempo necessario per completare i task[74].

3.3 Metodi per valutare il carico di lavoro mentale

Il carico di lavoro mentale (*mental workload* o *cognitive workload*) è una delle variabili più importanti nella psicologia, ergonomia e human factors. Esistono diverse definizioni di mental workload:

- "the mental effort that the human operator devotes to control or supervision relative to his capacity to expend mental effort? workload is never greater than unity."[21];
- "mental workload may be viewed as the difference between the capacities of the information processing system that are required for task performance to satisfy performance expectations and the capacity available at any given time."[28];
- "the cost of performing a task in terms of a reduction in the capacity to perform additional tasks that use the same processing resource."[41];
- "the relative capacity to respond, the emphasis is on predicting what the operator will be able to accomplish in the future."[56];
- "Mental workload refers to the portion of operator information processing capacity or resources that is actually required to meet system demands."[90].

Tutte queste definizioni possono essere riassunte nella definizione di carico di lavoro mentale come "costrutto mentale che riflette la tensione mentale derivante dall'esecuzione di un compito"[13].
Il *mental workload* può essere analizzato sotto due aspetti:

- **aspetto soggettivo** la cui validità ed affidabilità è dimostrata da test e questionari (NASA-TLX, OW, SWAT ecc);

- **aspetto fisiologico** registrato in fase di test che tuttavia può variare da soggetto a soggetto (analisi del battito del cuore, del movimento degli occhi e di altri aspetti psicosomatici).

L'usabilità, così come il carico di lavoro mentale, è uno dei costrutti dall'ergonomia; sviluppatisi di pari passo hanno mostrato, con il passare degli anni, una sempre maggiore correlazione. Per questo motivo molti studi hanno cercano di individuare il carico di lavoro mentale (o *mental wordload*) per valutare l'usabilità dei sistemi.
Inoltre spesso si fa riferimento a *mental wordload* come misuratore dell'efficienza umana.
Analizziamo ora alcune delle procedure più utilizzate per valutare il carico di lavoro di un sistema sull'utente.

3.3.1 L'OW

L'*Overall Workload* (OW) *scale*[32] è stata definita come

> "A single, 20-step bipolar scale is used to obtain this global rating. A score from 0 to 100 (assigned to the nearest 5) is obtained"[34]

L'OW, infatti, consiste in una scala unidimensionale da 0 a 100; ogni soggetto fornisce una valutazione in termini numerici, dove 0 rappresenta un carico di lavoro molto basso e 100 un carico di lavoro molto elevato.

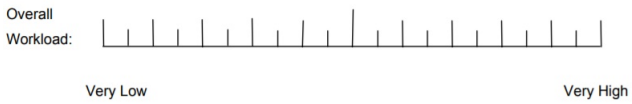


Figura 3.1: struttura dell'OW

L'OW rappresenta uno strumento che richiede poco tempo per essere completata e di facile utilizzo; non solo: richiede poco tempo per imparare come amministrarlo, prepararlo per la somministrazione ed analizzare i dati ottenuti.
Nel corso degli anni sono state fatte molte considerazioni a riguardo:

- è stato provato che l'OW produce risultati comparabili con il NASA-TLX[83];
- è stato sottolineato come l'OW sia un modo eccellente per verificare il carico di lavoro mentale attraverso una scala unidimensionale, considerandolo quasi alla pari a livello di sensibilità[16, 34];
- è stato definito utile come strumento di screening per identificare i potenziali punti critici del carico di lavoro[34].

Nonostante l'OW abbia molti punti di forza, convalidanti da studi nel corso degli anni, rimane uno strumento poco noto ed utilizzato.

3.3.2 Il NASA-TLX

Il *NASA Task Load Index* (NASA-TLX) è una scala multidimensionale progettata nel 1988[29] per ottenere stime del carico di lavoro durante l'esecuzione di un'attività o al suo termine.
Il NASA-TLX è costituito da sei sottoscale pesate:

1. richiesta mentali;
2. richiesta fisiche;
3. richiesta temporale;
4. prestazione;
5. frustrazione;
6. sforzo.

Le sottoscale sono individuate sotto l'assunzione che la loro combinazione, ponderata e pesata, fornisca la stima del carico di lavoro percepito per la

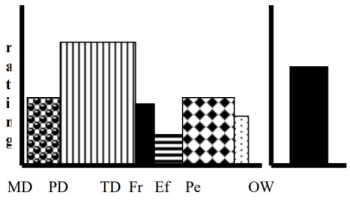


Figura 3.2: le sottoscale ed il valore complessivo del carico di lavoro

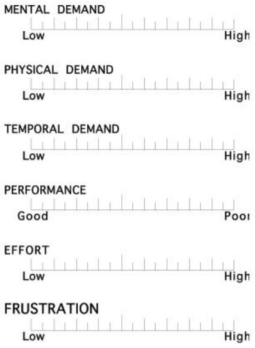


Figura 3.3: struttura del NASA-TLX

realizzazione di un task.
Dopo più di 20 anni di utilizzo, il NASA-TLX ha ottenuto una certa notorietà; viene utilizzato come benchmark per confrontare l'efficacia di altri strumenti, teorie o modelli; inoltre le continue valutazioni, modifiche, estensioni ed applicazioni a nuovi ambiti contribuiscono a renderlo sempre più attuale.
Nel corso degli anni, alcuni studi, hanno teorizzato come modificare il NASA-TLX[31]:

- aggiungendo sottoscale per meglio comprendere il carico mentale percepito;
- eliminando alcune delle sottoscale per semplificare il NASA-TLX;
- ridefinendo le sottoscale esistenti per migliorare la pertinenza rispetto ai task;

- eliminando del tutto il processo di ponderazione o pesando le sottoscale e poi analizzandole individualmente (questa versione viene definita *Raw TLX* tuttavia vi sono risultati contrastanti tra gli studi che cercando di valutarne la correlazione con il NASA-TLX [16, 33, 53])

3.3.3 Il SWAT

Il *Subjective Workload Assessment Technique* (SWAT), nato nei primi anni 80 ma formalizzato poi nel 1988[65], è uno strumento multidimensionale per valutare il *mental workload* basato sull'analisi di tre fattori:

1. **carico in termini di tempo** (T): la quantità di tempo libero disponibile nella pianificazione, esecuzione e monitoraggio di un task;
2. **carico mentale** (E): lo sforzo mentale e la pianificazione necessari per eseguire un task;
3. **carico in termini di stress psicologico** (S): il rischio, la confusione, la frustrazione e l'ansia associati all'esecuzione del task.

Per ognuno di questi livelli viene assegnato un valore (1 = basso, 2 = medio o 3 = alto) per poi ottenere un punteggio riepilogativo del carico di lavoro mentale.
A differenza degli strumenti già visti dove veniva utilizzata una scala Likert, in questo caso la valutazione è accompagnata da un testo descrittivo per aiutare l'utente nella fase di valutazione del task.
Lo SWAT è rappresentato da una procedura in due step:

- **setp 1 - scale development**: una procedura complessa e costosa, in termini di tempo e risorse, volta alla realizzazione della scala di valutazione:

    1. vengono rappresentate tutte le possibili combinazioni dei livelli delle tre dimensioni su 27 carte (la carta [T=1, E=1, S=1] rappresenterà il caso migliore mentre la carta [T=3, E=3, S=3] rappresenterà il caso in cui il carico di lavoro mentale è massimo e di conseguenza il caso peggiore);

I. Time Load
1. Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.
2. Occasionally have spare time. Interruptions or overlap among activities occur infrequently.
3. Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.
II. Mental Effort Load
1. Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention.
2. Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainly, unpredictability, or unfamiliarity. Considerable attention required.
3. Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.
III. Psychological Stress Load
1. Little confusion, risk, frustration, or anxiety exists and can be easily accommodated.
2. Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance.
3. High to very intense stress due to confusion, frustration, or anxiety. High extreme determination and self-control required.

Figura 3.4: struttura del SWAT

2. una volta realizzate le carte l'operatore le ordina in base alla classificazione che riflette la propria percezione di aumento del carico di lavoro;

3. un'analisi computerizzata genera, attraverso tecniche di misurazione e ridimensionamento, una classificazione a scala con un intervallo da 0 a 100; in questo modo ognuna delle 27 possibili configurazioni *T-E-S* ha un valore sulla scala degli intervalli.

- **setp 2 - event-scoring**: questa fase rappresenta la valutazione effettiva del *mental workload* di un task; è una fase molto più semplice e breve:

  1. viene chiesto agli utenti di eseguire specifici task e valutarli in relazione al carico di tempo, sforzo mentale e carico di stress scegliendo una delle risposte del test, si ottiene così per ogni partecipante un punteggio (ad esempio 2-3-1);

  2. sfruttando la scala ottenuta con lo **step 1** si riconduce ad ogni valutazione il valore della scala corrispondente come valutazione del carico di lavoro del task espresso in un punteggio numerico tra 0 e 100;

Lo SWAT rappresenta una novità rispetto agli altri strumenti di analisi del mental workload e, più in generale, dell'usabilità.
La sua peculiarità è che non utilizza una scala arbitraria bensì viene modellata attraverso la percezione umana. Hendy sostenne che

> "(gains) insight into the mechanism of human information processing resources, together with the notion that it is possible to derive a model, by some rational procedure, that has greater validity than that of an arbitrarily chosen model"[33]

facendo riferimento a come, grazie al metodo utilizzato per ottenere la scala di valutazione, si ottenga una maggiore validità del modello rispetto ad una scala arbitrariamente scelta.
Nel corso degli anni sono state fatte diverse considerazioni riguardo al SWAT:

- è stato riscontrato che le tre dimensioni "mancano di ortogonalità soggettiva"[12] sostenendo che alti livelli di carico di tempo aumentano di conseguenza il livello di carico di lavoro mentale;
- è stato osservato che il 43% degli utenti esperti che segue la procedura per la realizzazione della scala SWAT fallisce al primo tentativo di ordinamento delle carte per la creazione della scala, ciò implica che utenti inesperti potrebbero incorrere in un tasso ancora maggiore di fallimento[34];
- alcuni articoli hanno individuato che il NASA-TLX, confrontato con la scala SWAT, è generalmente considerata la scala migliore per misurare il carico di lavoro mentale[34, 63];
- Wierwille e Eggemeier hanno enunciato che la SWAT è potenzialmente in grado di identificare "meccanismi cognitivi che influenzano il carico di lavoro mentale"[88];
- alcuni studi indicano che la scala SWAT si rivela utile per stimare i cambiamenti nel carico di lavoro mentale[88, 23, 19];

- Colle e Reid hanno scoperto che la scala SWAT è più sensibile ai cambiamenti di difficoltà e contesto rispetto al NASA-TLX[19].

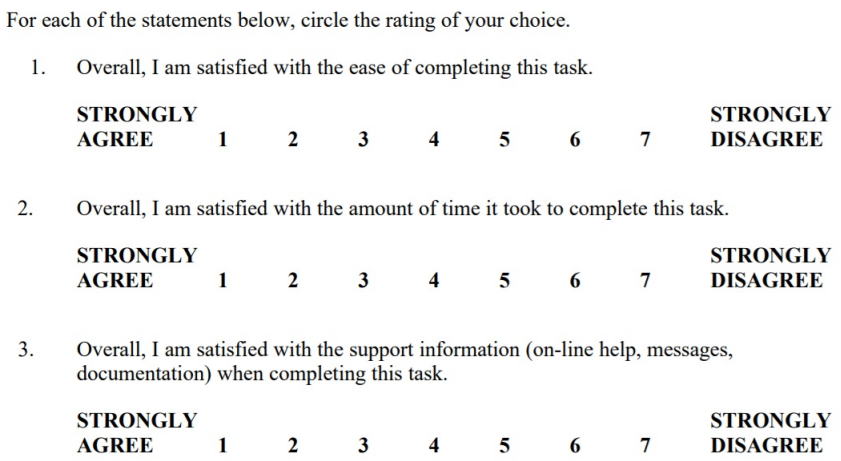
Analizzando la letteratura che fa riferimento al SWAT è evidente come la sua natura multidimensionale e la scala creata a partire dalla percezione degli utenti fornisca un punto di vista molto interessante.
Il risvolto della medaglia è però rappresentato dal fatto che sia molto costoso in termini di tempo e risorse e sopratutto soggetto ad errori in fase di preparazione. Queste motivazioni lo rendono, sebbene valido, raramente utilizzato a discapito del ben più noto e semplice NASA-TLX.

3.4 Metodi per valutare la soddisfazione

Alcuni degli strumenti utilizzati per valutare un sistema si basano sulla soddisfazione percepita dagli utenti. Si tratta di test post-task o post-scenario volti a valutare la capacità di completare i task e la percezione avuta dagli utenti a riguardo.

3.4.1 L'ASQ

L'*After-Scenario Questionnaire* (ASQ)[45], utilizzato dall'IBM, è un questionario costituito da tre quesiti formati ognuno da una scala Likert di sette punti.
Essendo un questionario post-scenario viene sottoposto ai partecipanti ad un test dopo il completamento di ogni task.
L'ASQ valutata: la facilità di completamento dell'attività, il tempo per completare un'attività e l'adeguatezza delle di informazioni di supporto (guida in linea, messaggi e documentazione).
La correlazione tra i punteggi dell'ASQ e il fallimento o il successo dello scenario (codificato come 0 = fallimento e 1 = successo) ha dimostrato che i partecipanti che completano con successo uno scenario tendono a dare un voto minore (più favorevole) provando la validità dell'ASQ. Inoltre l'analisi



For each of the statements below, circle the rating of your choice.

1. Overall, I am satisfied with the ease of completing this task.

STRONGLY AGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**STRONGLY DISAGREE**

2. Overall, I am satisfied with the amount of time it took to complete this task.

STRONGLY AGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**STRONGLY DISAGREE**

3. Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing this task.

STRONGLY AGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**STRONGLY DISAGREE**

Figura 3.5: struttura dell'ASQ

della varianza dell'ASQ dimostra che il punteggio ottenuto è ragionevolmente sensibile.
I primi studi pubblicati sull'ASQ coinvolgevano un campione di 48 utenti sebbene il numero ideale sarebbe 120 (5 partecipanti x 8 scenari x 3 items/scenario), rendendo questo test costoso sia in termini di risorse necessarie. Rimane tuttavia un test affidabile, valido e sensibile.

3.4.2 Il CSUQ

Il *Computer System Usability Questionnaire* è uno strumento sviluppato all'interno della IBM da Lewis[49]. É un questionario after-scenario composto da 19 quesiti la cui risposta è rappresentata da una scala Likert di 7 elementi (più la possibilità di non dare risposta alla domanda selezionando NA), somministrato per la prima volta ai dipendenti della IBM per testare un sistema di mailing.
Creato a partire dal PSSUQ (*Post-Study System Usability Questionnaire*), modificato opportunamente, rappresenta un'interessante ed efficace strumento per la valutazione della soddisfazione degli utenti.
Nel corso degli anni sono state elaborate versioni del CSUQ con 16 e 18 que-

siti. La peculiarità di questo questionario è che il numero di quesiti permette di valutare 3 sottoscale che forniscono una ripartizione più dettagliata dei diversi fattori che influenzano il sistema in analisi.

Overall Reaction to the Website		1	2	3	4	5	6	7		NA
1. Overall, I am satisfied with how easy it is to use this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
2. It was simple to use this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
3. I can effectively complete my work using this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
4. I am able to complete my work quickly using this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
5. I am able to efficiently complete my work using this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
6. I feel comfortable using this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
7. It was easy to learn to use this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
8. I believe I became productive quickly using this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
9. The website gives error messages that clearly tell me how to fix problems	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
10. Whenever I make a mistake using the website, I recover easily and quickly	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
11. The information (such as online help, on-page messages, and other documentation) provided with this website is clear	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
12. It is easy to find the information I need	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
13. The information provided by the website is easy to understand	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
14. The information is effective in helping me complete the tasks and scenarios	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
15. The organization of information on the website pages is clear	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
16. The interface of this website is pleasant	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
17. I like using the interface of this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
18. This website has all the functions and capabilities I expect it to have	strongly disagree	○	○	○	○	○	○	○	strongly agree	○
19. Overall, I am satisfied with this website	strongly disagree	○	○	○	○	○	○	○	strongly agree	○

Figura 3.6: struttura dell'CSUQ

Più nel dettaglio, nel CSUQ a 16 quesiti, è possibile individuare[51]:

- il **punteggio complessivo**: i punteggi medi delle domande da 1 a 16;
- l'**utilità del sistema**: i punteggi medi delle domande da 1 a 6;
- la **qualità dell'informazione**: i punteggi medi delle domande da 7 a 12;

- la **qualità dell'interfaccia**: i punteggi medi delle domande da 13 a 15.

3.4.3 La SUS

Il *System Usability Scale* (SUS) è stato definito da Brooke nel 1996[14] come

> "a quick and dirty tool to measure the perceived usability of a system"

ovvero uno strumento basilare per catturare in modo semplice l'usabilità percepita di un sistema.
Si tratta di un questionario post-study costituito da 10 quesiti le cui risposte sono rappresentate da una scala Likert di 5 elementi.
Il risultato della SUS è un punteggio nel range 0-100 che permette non solo la valutazione dell'usabilità di un sistema ma anche il suo confronto con altri sistemi.
Nato come uno strumento grezzo per l'analisi dell'usabilità con il passare degli anni la SUS è diventata uno degli strumenti più utilizzati.
Nei primi anni di vita della SUS sono stati condotti molti studi per verificarne la solidità e l'effettiva affidabilità dei risultati; negli anni non solo si è dimostrata solida ma ha anche più breve ed efficace di altri strumenti esistenti:

- nel 1993 Holyer[36] dimostrò che la SUS è correlata al 0,86 con il *50-item Software Usability Measurement Inventor* (Kirakowski, 1992), dimostrando che, anche riducendo drasticamente il numero di quesiti, il risultato complessivo rimane veritiero ed affidabile;

- nel 2004 Tullis e Stetson[84] hanno scoperto che la SUS supera il *Questionnaire for User Interface Satisfaction*[18] e il *Computer System Usability Questionnaire*[49] nel valutare l'usabilità dei siti web.

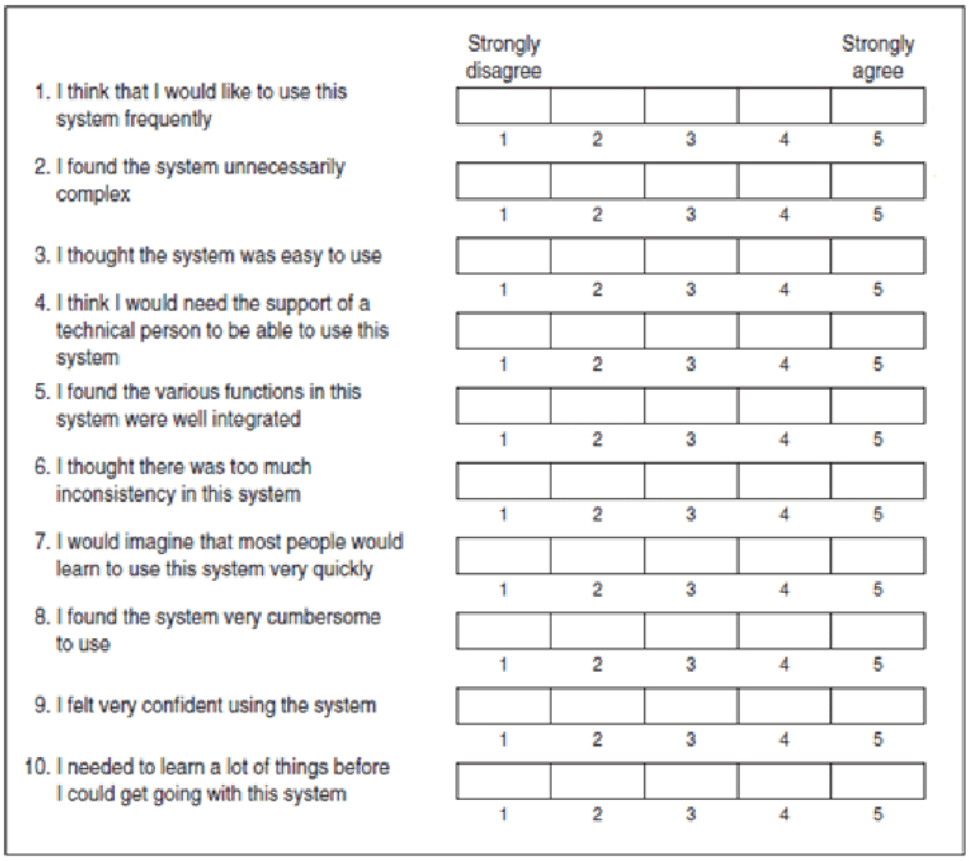


Figura 3.7: struttura della SUS

Brooke stesso, dopo circa 20 anni dalla pubblicazione dell'articolo di presentazione della SUS, ha pubblicato uno studio in cui venivano mostrati i risultati ottenuti in quegli anni, frutto dell'analisi di migliaia di utilizzi della SUS[15]. Infatti non solo la SUS negli anni è stata ampiamente utilizzata ma molti sono stati gli approfondimenti e le considerazioni fatte dagli studiosi. In particolare:

- nel 2004 Tullis e Stetson[84] realizzarono uno studio che provava che la SUS risulta affidabile anche con un campione molto ridotto;
- nel 2009 Bangor, Kortum e Miller[6] svolsero degli studi per poter trasformare il punteggio finale della SUS in una valutazione dalla A alla F per comprendere il significato assoluto del punteggio finale, una

novità dato che negli anni precedenti il punteggio era semplicemente considerato come percentuale di usabilità;

- nel 2009 Sauro e Lewis[52] ed anche Borsci, Federici e Lauriola[11] analizzarono una possibile bidimensionalità della SUS che, secondo alcuni studi, permetteva non solo di valutare l'usabilità di un sistema ma di valutarne anche l'apprendibilità.

Analizzando la letteratura attuale si può notare come la SUS, per quanto inizialmente ritenuto uno strumento *veloce e grezzo*, è stata adottata come misura standard di usabilità grazie alle sue caratteristiche prestazionali dimostrate, oltre ad essere libera e relativamente compatta.

3.4.4 L'UMUX

L'*Usability Metric for User Experience* (UMUX) è uno strumento sviluppato nel 2010 da Kraig Finstad[27] e i suoi colleghi della Intel con lo scopo di offrire un'alternativa più breve alla SUS.
L'UMUX si basa su alcune critiche smosse contro la SUS, ovvero:

- nel 1993 Diefenbach[24] enunciò che le scale Likert a sette punti superano le scale a cinque punti in termini di affidabilità, precisione e facilità d'uso (revisionando lo studio di Cox del 1980[20] in cui stabiliva che il numero ottimale opzioni delle scale Likert è sette);

- Finstad nel 2006[25] analizzò come nel quesito 8 della SUS la parola "*cumbersome*" risultasse ostica per gli inglesi non-nativi rendendo il risultato del test influenzabile dalla provenienza degli utenti a cui è sottoposto;

- Finstad[26] ha rilevato che gli utenti a cui vengono sottoposti test di usabilità sono più propensi a fornire interpolazioni non intere (ad es. "tre e mezzo" invece di "tre" o "quattro") rendendo così una scala Likert a sette punti più rappresentativa rispetto ad una a 5 punti (come quella della SUS);

- un questionario di 10 domande per valutare la sola usabilità può essere troppo costoso in termini di risorse, sopratutto se è necessario valutare altri aspetti del prodotto in analisi e, di conseguenza, bisogna somministrare ulteriori test oltre alla SUS;

- la SUS non permette di mappare concretamente efficacia, efficienza e soddisfazione di un sistema ma la sua usabilità in modo generale rendendolo uno strumento non diagnostico.

Il risultato delle analisi di Kraig Finstad ed il suo team è un questionario costituito da 4 domande le cui risposte sono formulate in una scala Likert a sette punti.

	The Usability Metric for User Experience Version 1		Strongly Disagree						Strongly Agree
			1	2	3	4	5	6	7
1	This system's capabilities meet my requirements.		o	o	o	o	o	o	o
2	Using this system is a frustrating experience.		o	o	o	o	o	o	o
3	This system is easy to use.		o	o	o	o	o	o	o
4	I have to spend too much time correcting things with this system.		o	o	o	o	o	o	o

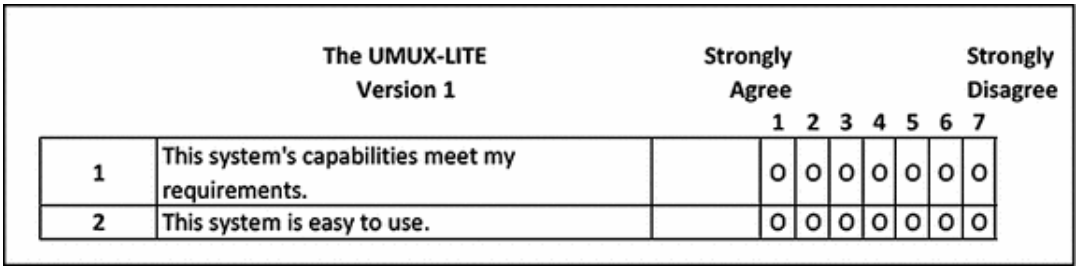
Figura 3.8: struttura dell'UMUX

Analizzando la correlazione tra SUS e UMUX è stato osservato che vi è un tasso di correlazione superiore al 0,8.
Nonostante i presupposti di partenza, tuttavia, l'UMUX misura l'usabilità percepita di un sistema, senza fornire ulteriori informazioni riguardo efficienza, efficacia, errori ecc.
L'UMUX rappresenta quindi un'alternativa affidabile, valida e sensibile alla SUS visti i risultati ottenuti ed il buon grado di correlazione nonostante il numero molto minore di quesiti. Nonostante la sua compattezza ed i buoni risultati ottenuti l'UMUX sembra non essere riuscito a superare la SUS in termine di utilizzo. É infatti lampante, cercando in rete, come la fama della SUS surclassi quella dell'UMUX. Bisogna tuttavia tener conto che la SUS è

nato quasi 20 anni prima rendendolo una pietra miliare dell'usabilità difficile da scalfire.

3.4.5 L'UMUX-LITE

Nel 2013 Lewis, Utesch e Maher[44] pubblicarono l'UMUX-LITE, una versione ridotta dell'UMUX composto da due soli quesiti le cui risposte sono formulate in scale Likert di sette elementi. L'UMUX-LITE mostra una buo-



	The UMUX-LITE Version 1		Strongly Agree 1	2	3	4	5	6	Strongly Disagree 7
1	This system's capabilities meet my requirements.		O	O	O	O	O	O	O
2	This system is easy to use.		O	O	O	O	O	O	O

Figura 3.9: struttura dell'UMUX-LITE

na correlazione, superiore al 0.8, sia con l'UMUX sia con la SUS, rendendolo uno strumento molto interessante per effettuare test d'usabilità percepita. L'UMUX-LITE, nonostante la sua compattezza e affidabilità, è tutt'ora poco noto ed utilizzato; viene prediletto l'uso della SUS o, al più, quello dell'UMUX nonostante le sue caratteristiche dovrebbero renderlo più appetibile a chi vuole effettuare un test d'usabilità veloce e valido.

Capitolo 4

Statistiche di utilizzo degli indicatori di usabilità

4.1 Analisi di singole tecniche

Per analizzare la fama e la mole di utilizzo delle tecniche per misurare l'usabilità, inteso come quantità di fonti che citano le stesse, è stata effettuata una ricerca per parole chiave su Google e Google Scholar, in modo tale da individuare quante risorse online citano gli strumenti descritti nel capitolo precedente.
Dalla ricerca è stato escluso l'OW dato che sia la sigla, sia il nome completo dello strumento (overral workload) creavano risultati falsati.
Per alcuni degli strumenti è stato aggiunto il suffisso "usability" per evitare di incorrere in risultati non inerenti (ad esempio cercando SWAT su Google si ottengono 106.000.000 risultati ma la maggior parte fa riferimento al corpo speciale americano). Inoltre, ovviamente, sono state usate le virgolette (") per incapsulare i nomi completi delle tecniche.
Dato che i vari strumenti per misurare l'usabilità sono stati sviluppati nel corso degli anni è stata fatta una ricerca generale (solo ricerca per parole chiave senza applicare filtri) ed una che tiene conto solo delle risorse presenti successive al 2010. In questo modo è stato possibile catturare sia il livello di

utilizzo e studio generale, sia quello ponderato agli ultimi 10 anni di applicazione.
Va tenuto in conto che il web è estremamente volatile e quindi effettuando le stesse ricerche, anche ad un singolo giorno di distanza, si possono ottenere risultati leggermente diversi.

	generale		dal 2010	
parole chiave	Google	Scholar	Google	Scholar
SUS usability	3.130.000	65.500	605.000	29.500
NASA-TLX	931.000	27.000	64.300	17.000
system usability scale	194.000	22.800	32.000	18.000
NASA task load index	62.100	14.200	14.100	11.600
ASQ usability	366.000	3.830	8.060	2.930
SWAT usability	1.950.000	3.730	18.200	2.790
subjective workload assessment technique	13.700	2.520	1.730	1.520
CSUQ	69.800	1.830	8.470	1.520
computer system usability questionnaire	9.230	1.380	1.570	1.170
UMUX	96.700	1.320	8.940	926
after-scenario questionnaire	8.000	950	1.190	795
usability metric for user experience	3.530	607	945	597
UMUX-LITE	14.900	367	1.840	361

Tabella 4.1: Risultati ottenuti dalla ricerca per parole chiave, ordinati in ordine decrescente in base ai risultati generali di Google Scholar

Analizzando i risultati ottenuti dalla ricerca risulta evidente che gli strumenti più citati sono la SUS ed il NASA-TLX seguiti a grande distanza da UMUX e SWAT.
La ricerca per "SWAT" genera un gran numero di risorse come risultato in quanto l'aggiunta del suffisso "usability" riesce ad eliminare i vari riferimenti alla *Special Weapons And Tactics* ma intercetta la *Samba Web Administration Tool*, un'applicazione web per la configurazione di Linux rendendo il

dato falsato.
Il valore risultante dalla ricerca di "UMUX" mostra un certa rilevanza del questionario, tuttavia quest'ultima non è confermata da Google Scholar.
É interessante notare come, dal 2010 ad oggi, nonostante l'UMUX e l'UMUX-LITE abbiano un numero minore di quesiti e una buona correlazione con la SUS rimangono poco usati.
É evidente quindi che non sempre rendere più breve e coinciso un test lo renda anche più utilizzato e apprezzato.

4.2 Utilizzo combinato di più metodi

Dopo aver analizzato le risorse presenti in rete per i singoli metodi è stata effettuata una seconda ricerca.
In questo frangente sono stati combinati, a coppie, i diversi strumenti per individuare quale combinazioni di tecniche è più citata.
Molto spesso, infatti, non si utilizza uno solo degli strumenti ma lo si usa in abbinamento ad uno o più strumenti per valutare l'accuratezza dei risultati ottenuti e fare nuove considerazione tra correlazione e dipendenza delle metriche dell'usabilità.

Analizzando i risultati della ricerca emerge che la SUS ed il NASA-TLX sono le tecniche più spesso utilizzate congiuntamente per l'analisi dell'usabilità dei sistemi.
Questo risultato era già emerso parzialmente dalla ricerca precedente tuttavia la ricerca appena effettuata conferma come spesso si analizzi, per verificare l'usabilità di un sistema, non solo la soddisfazione percepita dall'utente ma anche il carico di lavoro mentale.

parole chiave	Google	Scholar
SUS NASA-TLX	77.300	3.230
NASA-TLX SWAT	51.000	2.850
SUS UMUX	17.200	1.370
SUS CSUQ	16.200	902
SUS ASQ	9.330	706
SUS UMUX-LITE	12.900	346
ASQ CSUQ	4.830	297
SUS SWAT	454.000	266
UMUX CSUQ	1.870	202
CSUQ NASA-TLX	1.400	137
ASQ NASA-TLX	45.200	114
UMUX ASQ	31.000	98
UMUX-LITE CSUQ	1.150	82
UMUX NASA-TLX	1.380	50
UMUX-LITE ASQ	830	32
UMUX SWAT	1.600	28
UMUX-LITE NASA-TLX	1.330	25
ASQ SWAT	83.300	16
CSUQ SWAT	324	16
UMUX-LITE SWAT	855	2

Tabella 4.2: Risultati ottenuti dalla combinazione delle tecniche

Capitolo 5

Nuovi approcci

5.1 I test di usabilità oggi

Nelle sezioni precedenti sono state presentate le tecniche per valutare l'usabilità più consolidate ed utilizzate.
Per ognuna delle metriche esistono molteplici strumenti che permettono di ottenere una valutazione ed visione più generale del sistema che si sta valutando. Molto spesso si utilizzano più strumenti, che misurano metriche diverse, per valutare più aspetti del sistema ed avere quindi una panoramica su vari aspetti.
Come visto dalle statistiche del capitolo precedente la SUS ed il NASA-TLX, unitamente o singolarmente, sono gli strumenti più utilizzati.
Emerge quindi come l'usabilità generale dei sistemi ed il carico mentale rappresentano due delle metriche più analizzate per quanto riguarda l'usabilità e la user experience.
É tuttavia singolare, sotto un certo punto di vista, come due strumenti piuttosto datati vengano preferiti a nuovi, solidi e più brevi tecniche.
Ne è un esempio l'UMUX, nato proprio a partire della SUS e con un ottima correlazione con la stessa, che con le sue 4 domande è molto più snello e veloce da somministrare ma rimane poco utilizzato.
Queste considerazioni sorgono in modo automatico vista la percezione del-

l'usabilità attuale.
Mentre agli inizi del 1990 i prodotti di cui si voleva valutare l'usabilità erano relativamente pochi e quindi si era più propensi a spendere tempo e risorse per una valutazione a tuttotondo oggi non è più così.
Negli anni '90 gli strumenti per la valutazione dell'usabilità erano rappresentati da questionari di tra i 50 ed i 10 quesiti, che permettevano di produrre un risultato generico ed inoltre di avevano una caratteristica diagnostica; lo scopo dei test era quindi non solo di individuare la presenza di problemi di usabilità ma anche di cercare di trovare in modo puntuale le cause di questi problemi, anche a discapito di test più onerosi in termini di risorse.
Nell'era delle app, siti web e start-up che nascono ad una velocità inimmaginabile, il processo di realizzazione e messa sul mercato è molto più breve e più esigue sono le risorse a disposizione e, di conseguenza, il tempo ed il denaro per eseguire test d'usabilità.
Nasce così la necessità di rendere i test d'usabilità più compatti e semplici da somministrare.

5.2 Adeguamento delle metodologie alle nuove esigenze

A partire dagli anni '90 sempre più sforzi sono stati fatti per sviluppare tecniche per l'analisi dell'usabilità sempre più coincisi.
In alcuni casi si tratta della formalizzazione di nuove tecniche, in altri della rivalutazione di tecniche già teorizzate negli anni precedenti ma poco utilizzate.
La peculiarità di questi strumenti è che sono tutti costituiti da un solo quesito mantenendo una buona correlazioni con le tecniche più consolidate e formate da più quesiti.

5.2.1 SMEQ

Il *Subjective Mental Effort Question* (SMEQ) teorizzato nel 1985[92] e poi ripreso nel 1994 da Houwing[37] è uno strumento per misurare lo sforzo mentale che le persone percepiscono contestualmente all'esecuzione di un task.
É rappresentato da una scala con nove etichette (da "per nulla difficile da fare" a "tremendamente difficile da fare") situate su una linea verticale millimetrata con valori ad intervallo di 10 che vanno da 0 a 150.
É possibile compilarlo, nella versione cartacea, tracciando un segno sulla linea dove si ritiene che si collochi lo sforzo percepito nell'esecuzione del task. Sauro e Dumas[71] svilupparono una versione online dove la linea è rappresentata da uno slider che gli utenti possono spostare nella posizione che ritengono più rappresentativa.

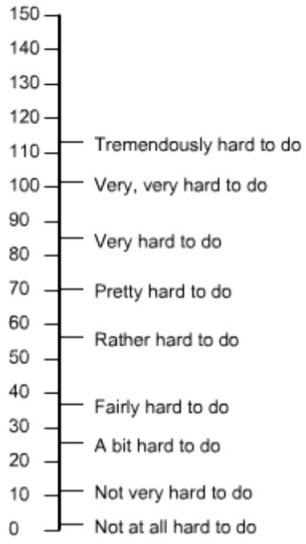


Figura 5.1: struttura del SMEQ

Il SMEQ rappresenta uno strumento solido e veloce da somministrare, tuttavia alcune critiche sono state smosse riguardo la difficoltà di comprensione della scala di valutazione che, non essendo uniforme, potrebbe portare inizialmente un senso di confusione nei partecipanti e nei somministratori del

test.
Secondo Sauro e Dumas[71] il SMEQ ha una significante correlazione con la SUS in termini di punteggio, tempo di completamento, tasso di completamento ed errori. Ciò dimostra la validità di questa tecnica.
Confrontando il SMEQ con con l'UME (che verrà analizzato nella prossima sezione) risulta molto più semplice da comprendere dai partecipanti dei test e, per questo motivo, prediletto tra i due.

5.2.2 UME

L'*Usability Magnitude Estimation* (UME)[67], a differenza delle tecniche precedentemente analizzate, ha una peculiarità particolare: i partecipanti al test creano la propria scala di valutazione.
Dopo l'esecuzione dei task agli utenti è chiesto di esprimere una valutazione numerica maggiore di 0 in riferimento alla complessità del task.
Tale valutazione, e di conseguenza il punteggio, servono per costituire le basi delle scale di valutazione.
Per ogni task, quindi, ogni partecipane esprime la sua votazione (tenendo conto delle valutazioni fornite per i task precedenti) e, una volta raccolti tutti i dati, si realizza una scala che riassume le varie valutazioni soggettive.

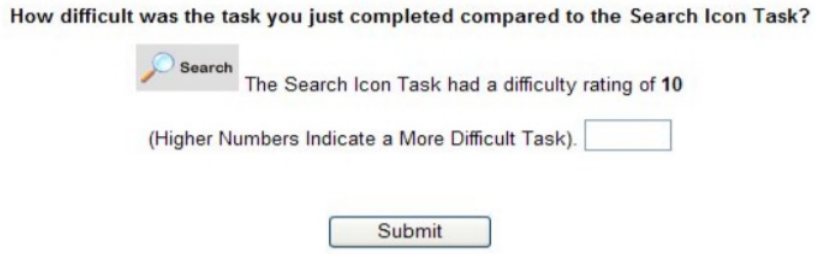


Figura 5.2: struttura dell'UME

L'UME è stato realizzato in risposta ad alcune critiche riguardanti le scale Likert, ritenute troppo riduttive in alcuni casi.
Nell'UME infatti ognuno può dare la propria valutazione senza limitazioni

(si può dare un punteggio da 1 a 5, da 1 a 100 o anche da 1 a 1000000 in base a come ogni singola persona percepisce la difficoltà).
Ovviamente ricondurre i punteggi ottenuti dall'UME a un valore confrontabile è più oneroso rispetto all'utilizzo di scale Likert.
Alcuni ricercatori[81] provarono ad inserire una condizione aggiuntiva all'UME per rendere i risultati più comprensibili e facilmente confrontabili: il valore assegnato doveva essere indicato in una scala da 1 e 100.
Questo studio non ottenne gradi risultati, al contrario, i partecipanti percepivano questa scala come una scala Likert ignorando e non comprendendo come dare giudizi in modo coerente con l'impostazione del metodo.
L'UME ha mostrato una buona correlazione con test composti da una scala Likert per tutti gli aspetti (risultato generale, errori) ma non per il tempo di esecuzione.

5.2.3 SEQ

Il *Single Ease Question* (SEQ) è un questionario definito da Sauro e Doumas[71] costituito da una sola, unica domanda rappresentata da una scala Likert di 7 punti.
Jeff Sauro stesso enuncia che rispondere al questionario e rispondere alla domanda

> "How easy or difficult was this task?"

sono equivalenti.

Figura 5.3: struttura del SEQ

Il SEQ è stato sviluppato per rispettare non solo le proprietà psicometriche dei questionari (affidabilità, sensibilità e validità) ma anche con lo scopo di

realizzare uno strumento breve da somministrare ed amministrare, a cui è facile rispondere e semplice da valutare.
Sauro nel 2018[73] pubblicò un articolo per validare ulteriormente il SEQ. Nell'articolo il SEQ non viene mai nominato, il fulcro è invece la SUS ed il numero di quesiti che la compongono.
Nell'articolo quello che Sauro fa è analizzare il modificarsi dell'affidabilità e della validità della SUS al diminuire del numero di quesiti proposti.
Emerge così che il solo quesito numero 3, somministrato da solo, fornisce un'accuratezza dei risultati del 90% rispetto al test somministrato per intero.
Altra conclusione degna di nota è che, secondo le analisi effettuate dal ricercatore, con soli due quesiti (il 3 e l'8) si ottiene il 96% dell'accuratezza e dal terzo elemento in poi i miglioramenti sono effimeri e di poca considerazione.

Figura 5.4: quesito numero 3 della SUS

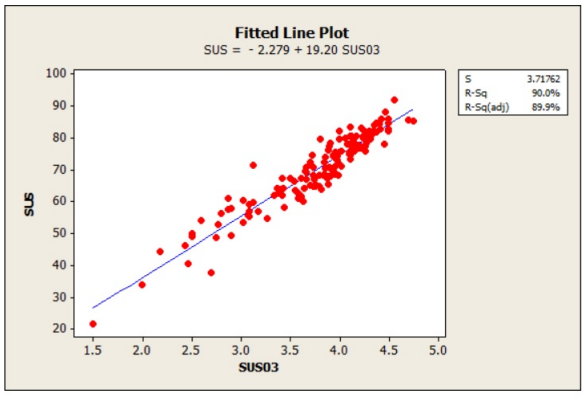


Figura 5.5: correlazione tra il risultato della SUS e delle risposte al quesito numero 3

Non solo Sauro nel corso degli anni ha messo in discussione la SUS per quan-

to riguarda il numero di quesiti, infatti nel 2009, alcuni studiosi guidati da Bangor[6] enunciarono una tecnica per trasformare il punteggio della SUS in un "voto" parlante e confrontabile, notando come un singolo quesito sembrava riassumere l'intera SUS con un certo grado di correlazione.
Per realizzare una scala di valutazione della SUS, Bankor et al, aggiunsero alla SUS un quesito la cui risposta era rappresentata da una scala Likert di 7 elementi (da "worst imaginabile" a "best imaginable").

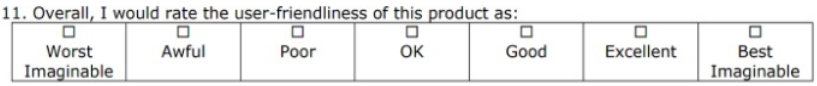


Figura 5.6: quesito aggiunto alla SUS da Bankor, Kortum e Miller

Confrontando il risultato della SUS con la risposta alla domanda aggiuntiva, non solo individuarono una scala di valutazione per il questionario ma notarono anche come le risposte fornite a quella sola domanda sembravano riassumere in modo piuttosto esaustivo il risultato finale della SUS.
Dalle fonti citate sopra, unite alla buone prestazioni dimostrate nei test, il SEQ sembrerebbe avere tutte le carte in regola per sostituire la SUS nell'ottica di effettuare test d'usabilità facili e veloci da somministrare.

5.2.4 Dati a confronto

Analizziamo ora le citazioni e la presenza in rete di SEQ, UME, SMEQ a confronto con SUS e NASA-TLX per poter ottenere una valutazione più generale.

Il SEQ, l'UME e lo SMEQ, analizzando i dati della tabella, sono noti ed utilizzati seppure non allo stesso livello della SUS e NASA-TLX.
Il SEQ sembra avere una "marcia in più" rispetto agli altri due, quasi eguagliando i numeri del NASA-TLX per quanto riguarda il periodo dal 2010 ad oggi.

	generale		dal 2010	
parole chiave	Google	Scholar	Google	Scholar
SUS usability	3.130.000	65.500	605.000	29.500
NASA-TLX	931.000	27.000	64.300	17.000
SEQ usability	457.000	21.200	80.600	15.300
UME usability	591.000	10.800	17.100	6.710
SMEQ usability	3.530	440	302	357

Tabella 5.1: Risultati ottenuti dalla ricerca per parole chiave dei nuovi misuratori dell'usabilità a confronto SUS e NASA-TLX

Tuttavia questi risultati si ottengono cercando per parole chiave. Se invece, come fatto per raccogliere la documentazione per questa dissertazione, si ricercano parole generiche come "usability metrics" o "usability methods" raramente appaiono: quasi tutti gli articoli più attuali che cercano di valutare l'usabilità di nuovi sistemi sfruttano la SUS, il NASA-TLX e l'UMUX. Questo accade soprattutto per quanto riguarda articoli accademici presenti su Google Scholar, su Google invece è molto più facile trovare riferimenti a questi strumenti sopratutto nei siti che spiegano come eseguire un test d'usabilità in modo veloce, efficiente ed alla portata di tutti.
Emerge così come nell'ambiente didattico si rimanga più fedeli agli strumenti più accreditati mentre nel web spesso si affiancano ai più conclamati metodi quelli più accessibili grazie alla loro semplicità e velocità di applicazione.

5.3 Riflessioni sui dati osservati

I dati raccolti nella sezione precedente rappresentano un ottimo punto da cui prendere spunto.
Abbiamo a disposizione degli strumenti semplici e, sopratutto, brevi da somministrare, perché non sfruttarli al massimo delle loro potenzialità?
Come visto il SEQ, così come altri indicatori, hanno una buona correlazione con la SUS, il che fa presupporre che è possibile condurre uno studio d'usabi-

lità sfruttando un unico quesito i cui dati risultanti sono accreditati e solidi.
Una sola domanda, è chiedere tanto?
La risposta la si può trovare nella tecnologia attuale.
Moltissime piattaforme sfruttano pop-up o semplici schermate per interagire con l'utente.
Ne è un esempio YouTube che, di tanto in tanto, chiede all'utente di rispondere a un sondaggio per fini commerciali.

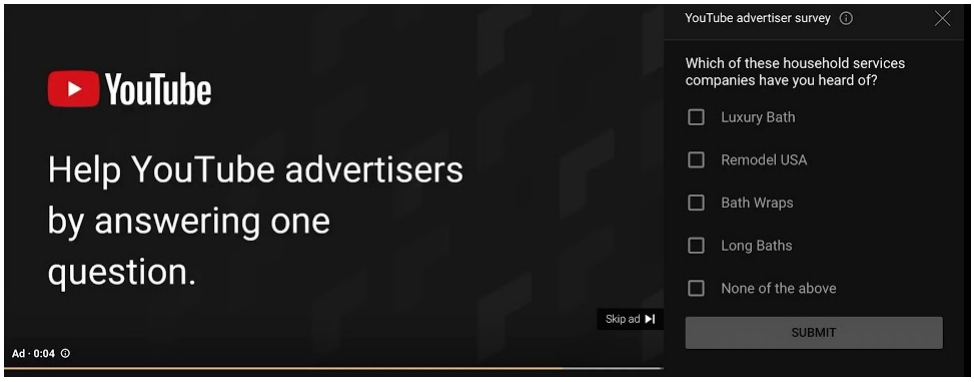


Figura 5.7: sondaggio a scopo pubblicitario di YouTube

Google Play chiede agli utenti di valutare le applicazioni presenti nello store lasciando una valutazione da 0 a 5 stelle e aggiungendo un commento opzionale.

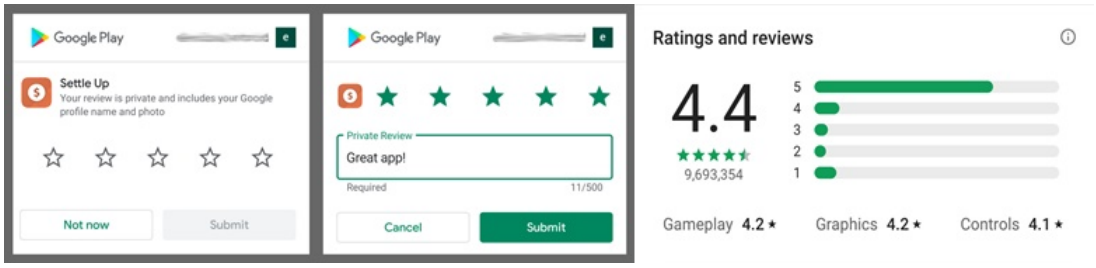


Figura 5.8: sistema di classificazione delle app su Google Play

Altro esempio di interazione con l'utente per fornire valutazioni è l'applicazione per smartphone UberEats (app per il food delivery) che dopo ogni ordine, una volta che l'utente ha ricevuto il cibo, chiede all'utente di fornire

una recensione sul rider, sul ristorante e sui singoli piatti orinati.

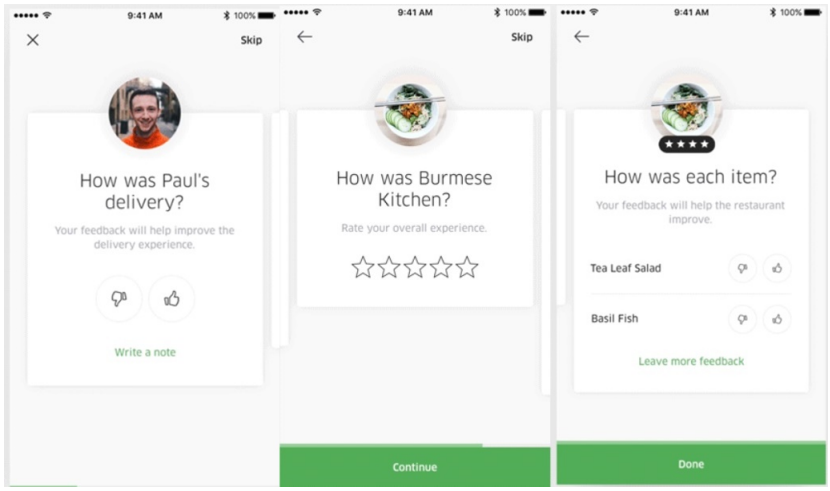


Figura 5.9: sistema di valutazione di UberEats

Questi solo solo alcuni degli esempi di piccole ma significative interazioni chieste all'utilizzo durante la navigazione web o in-app.
Emerge così la propensione degli sviluppatori a cercare il feedback degli utenti, sebbene per scopi pubblicitari o per la valutazione e creazione un sistema di ranking generale.
Parallelamente si osserva una propensione degli utenti a fornire feedback, sopratutto se si tratta solo di effettuare un paio di click.
Sorge spontaneo quindi chiedersi perché non utilizzare un sistema analogo per valutare l'usabilità dei sistemi?
A prescindere dalla possibilità o meno di effettuare dei test d'usabilità in fase di progettazione e realizzazione sarebbe molto interessante poi raccogliere dati da utenti effettivi, ottenendo così un campione molto alto a costo 0.
Gli scenari che si aprono in questo contesto sono molti: ad esempio si potrebbe chiedere all'utente come valuta l'usabilità del sistema (sotto l'aspetto della soddisfazione o secondo la/le metrica a cui si è più interessati) oppure si potrebbe ripetere questa somministrazione nel tempo e così via.
Sarebbe infatti interessante analizzare come la percezione dell'usabilità varia con il passare del tempo, man mano che vengono rilasciati aggiornamenti e,

in modo indiretto, in contrapposizioni a sistemi concorrenti.
Oltre a valutare l'usabilità generale del sistema si potrebbero aggiungere meccanismi di controllo che, durante il primo utilizzo o in momenti successivi, chiedono all'utente di valutare come l'attività appena svolta sia stata percepita.

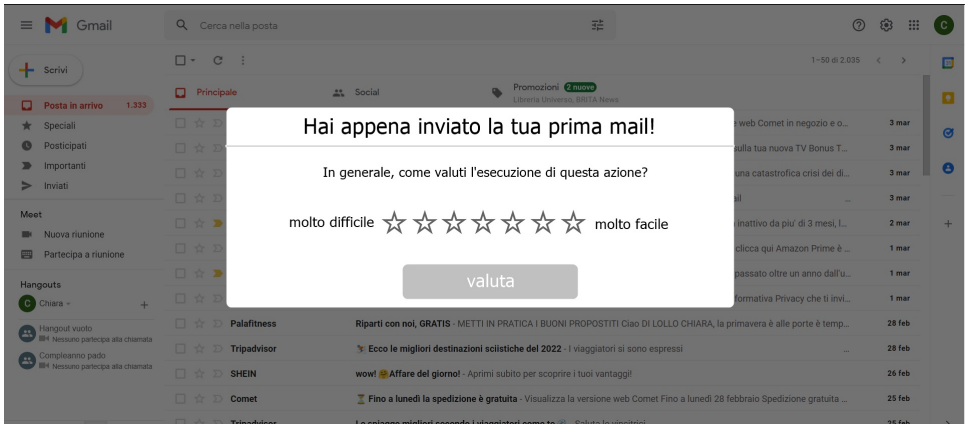


Figura 5.10: esempio di utilizzo del SEQ per la valutazione del task 'invio mail'

Questo aspetto è molto interessante a livello diagnostico in quanto permetterebbe di intervenire a livello di interfacce, procedure e, nei casi più complessi, a livello di documentazione.
In rete è difficile trovare informazioni riguardo a test d'usabilità eseguiti con cadenza più o meno regolare nel tempo. Probabilmente questo è dovuto al fatto che le aziende che raccolgono tali dati li mantengono privati per questioni concorrenziali o comunque come strategia di analisi e sviluppo interno. Data la scarsità dell'argomento in rete questo appare come un interessante punto di partenza per approcci futuri,

Conclusioni

La ricerca nell'ambito dell'usabilità nel corso degli anni è progredita in maniera esponenziale.
Da disciplina di nicchia ora rappresenta uno dei cardini della progettazione e programmazione.
Nel corso di questo elaborato sono stati portati alla luce le varie caratteristiche che la contraddistinguono, gli strumenti per misurarla e come essi si sono evoluti e sono mutati nel tempo.
Dalle analisi eseguite nei capitoli precedenti è emerso come gli strumenti più utilizzati sono il NASA-TLX e la SUS, metodi multidimensionali per la valutazione rispettiva di carico di lavoro mentale e soddisfazione.
Parallelamente emergono però indicatori come il SEQ che seppur composto da un solo quesito sembra essere all'altezza della SUS e fornire risultati altrettanto affidabili e validi.
Il SEQ delinea un punto di snodo molto importante: potrebbe essere la risposta alla mancanza di tempo e risorse per effettuare i test d'usabilità più classici è conosciuti.
Rappresenterebbe quindi l'escamotage per fare avvicinare al mondo dell'usabilità la nicchia di persone ancora scettiche e non disposte ad impiegare troppe risorse per un aspetto a volte ritenuto, erroneamente, trascurabile.
Sarebbe molto interessante, come sviluppo futuro, eseguire uno studio che metta a confronto il SEQ con la SUS ed ancora NASA-TLX ed OW.
Poter meglio osservare il grado di correlazione tra i metodi e valutare anche la variazione di affidabilità al variare del campione su cui lo studio è condotto

potrebbe rappresentare un interessante punto di vista per valutare se e quando utilizzare opportunamente il SEQ in sostituzione della SUS e viceversa ed anche quando utilizzare il NASA-TLX e quando utilizzare l'OW.
Quindi, partendo dal presupposto che la correlazione tra i metodi di misurazione è già stata verificata, proporre una guida completa che esplichi al meglio come utilizzare i due strumenti ed in quali circostanze, il tutto corredato da dati raccolti e valutazioni oggettive osservate in fase di test.
Come visto nel capitolo precedente attualmente molte applicazioni e servizi web propongono all'utente domande per scopi di marketing, pubblicitari, di ranking interno, di proliferazione ecc.
Proprio per questo motivo un altro interessante approccio sarebbe quello di implementare un sistema, anche basilare su cui effettuare dei test d'usabilità in fase di sviluppo (come SUS, CSUQ, UMUX, NASA, SWAT, ecc).
In seguito, una volta rilasciato al pubblico, sarebbe interessante somministrare all'utente alcuni brevi questionari, di quando in quando, per registrare ed analizzare non solo la percezione dell'utente ma anche come essa varia al variare del tempo.
Questo permetterebbe, ad un bassissimo costo (prettamente in fase di programmazione inizialmente e di analisi dei dati successivamente) di avere una panoramica completa di come gli utenti percepiscono il sistema.
Sarebbe possibile analizzare se l'esperienza dell'utente migliora, peggiora o rimane invariata del tempo ed, a seconda di questo dato sapere se e come intervenire.
Si potrebbe ottenere un feedback semi-immediato delle effetto degli aggiornamenti e delle migliorie apportate.
Infine, ma non meno importante, si potrebbe utilizzare i dati raccolti come dato di marketing per spazzare via la concorrenza.
Oggi, in un mondo in cui per ogni applicazione nuova ne vengono realizzate 4 equivalenti nel giro di poco, ciò che fa la differenza è il marketing e la capacità di distinguersi.
Proprio per questo a parità di prestazioni, funzionalità e design accattivante

quello che potrebbe fare la differenza è proprio l'usabilità.
Fu efficace per Steve Jobs nel 1984, oggi potrebbe essere il quid per fare la differenza.

Bibliografia

[1] M. Albers. Tapping as a measure of cognitive load and website usability. In *Proceedings of the 29th ACM international conference on Design of communication*, pages 25–32, 2011.

[2] Bill Albert and Tom Tullis. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. 2013.

[3] Ryan Alturki and Valerie Gay. Usability testing of fitness mobile application: Case study aded surat app. *International Journal of Computer Science & Information Technology (IJCSIT)*, 9, 2017.

[4] J.L. Alty. Can we measure usability? In *Proceedings of the Advanced Information Systems 1992 Conference*, pages 95–106, London: Learned Information, 1992.

[5] A Bangor, P.T. Kortum, and J.T. Miller. An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24:574–594, 2008.

[6] A. Bangor, P.T. Kortum, and J.T. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[7] M.I. Berkman and D. Karahoca. Re-assessing the usability metric for user experience (umux) scale. *Journal of Usability Studies*, 11(3):89–109, 2016.

[8] N. Bevan, J. Kirakowski, and J. Maissel. What is usability? In *Proceedings of the Fourth International Conference on Human Computer Interaction*, pages 651–655, 1991.

[9] Nigel Bevan and Miles Macleod. Usability measurement in context. *Behaviour and Information Technology*, 13:132–145, 1994.

[10] S. Borsci, S. Federici, M. Gnaldi, S. Bacci, and F. Bartolucci. Assessing user satisfaction in the era of user experience: An exploratory analysis of sus, umux and umux-lite. *International Journal of Human-Computer Interaction*, 31:484–495, 2015.

[11] S. Borsci, S. Federici, and M. Lauriola. On the dimensionality of the system usability scale: A test of alternative measurement models. *Cognitive Processes*, 10:193–197, 2009.

[12] S.P. Boyd. Assessing the validity of swat as a workload measurement instrument. In *Proceedings of the Human Factors Society - 27th Annual Meeting*, 1983.

[13] Cain Brad. A Review of the Mental Workload Literature . Standard, Defence Research and Development Canada Toronto, 2007.

[14] John Brooke. Sus: A "quick and dirty" usability scale. *Usability evaluation in industry*, pages 189–194, 1996.

[15] John Brooke. Sus: A retrospective. *Journal of Usability Studies*, 8(2):29–40, 2013.

[16] J.C. Byers, A.C. Bittner, and S.G. Hill. Traditional and raw task load index (tlx) correlations: Are paired comparisons necessary? *Advances in Industrial Ergonomics and Safety*, pages 481–485, 1989.

[17] J.C. Chen. Response-order effects in likert-type scales. *Educational and Psychological Measurement*, 51:531–540, 1991.

[18] J.P. Chien, V. Diehl, and K.L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of CHI '88 Conference on Human Factors in Computing Systems*, pages 213–218, 1988.

[19] H.A. Colle. Context effects in subjective mental workload rating. *Human Factors*, 40:591–600, 1998.

[20] E.P. Cox III. The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17:407–422, 1980.

[21] R. Curry and H. Jex. Mental Workload: Its theory and measurement . Standard, Final report of the control engineering group, 1979.

[22] D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13:319–339, 1989.

[23] D. De Waard. *The Measurement of Drivers' Mental Workload*. PhD thesis, University of Groningen, 1996.

[24] M.A. Diefenbach, N.D. Weinstein, and J. O'Reilly. Scales for assessing perceptions of health hazard susceptibility. *Health Education Research*, 8:181–192, 1993.

[25] K. Finstad. The system usability scale and non-native english speakers. *Journal of Usability Studies*, 1(4):185–188, 2006.

[26] K. Finstad. Response interpolation and scale sensitivity: evidence against five-point scales. *Journal of Usability Studies*, 5:104–110, 2010.

[27] K. Finstad. The usability metric for user experience. *Interacting with Computers*, 22:323–327, 2010.

[28] D. Gopher and E. Donchin. Workload - an examination of the concept. *Handbook of Perception and Human Performance*, 2:41–49, 1986.

[29] Sandra Hart and Lowell Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology*, 52:139–183, 1988.

[30] S.G. Hart. Theory and measurement of human workload. *Human productivity enhancement*, 1:396–455, 1986.

[31] S.G. Hart. Nasa-task load index (nasa-tlx): 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908, Los Angeles, CA, 2006.

[32] S.G. Hart, V. Battiste, and P.T. Lester. Popcorn: A supervisory control simulation for workload and performance research for workload and performance- research. In *Proceedings of the 20th Annual Conference on Manual Control*, pages 431–453, 1984.

[33] K.C. Hendy, K.M. Hamilton, and L.M. Landry. Measuring subjective workload: When is one scale better than many. *Human Factors*, 35:579–601, 1993.

[34] Susan Hill, Helene Iavecchia, James Byers, and Alvah Bittner. Comparison of four subjective workload rating scales. *The Journal of the Human Factors and Ergonomics Society*, 34:429–439, 1992.

[35] Andrew Hodrien and Terrence Fernando. A review of post-study and post-task subjective questionnaires to guide assessment of system usability. *Journal of Usability Studies*, 16:203–232, 2021.

[36] Andy Holyer. Methods for evaluating user interfaces. *Report number: 301 Affiliation: University of Sussex*, 1993.

[37] E.M. Houwing, M. Weithoff, and A. Arnold. Cognitive workload measurement. *Instructions and Background Materials*, 1994.

[38] Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability. Standard, International Organization for Standardization, 1998.

[39] Software enginnering - Product quality. Standard, International Organization for Standardization, 1991.

[40] Nielsen Jakob. *Usability Engineering*. 1993.

[41] A.F. Kramer and E.J. Sirevaag. A psychophysiological assessment of operator workload during simulated flight missions. *Human Factors*, 29:145–160, 1987.

[42] Riley Kundtz and Erinn Flandreau. Measuring Difficulty Doesn't Need to be Difficult: 4 Post-Task Metrics for a Better User Experience. `https://www.bentley.edu/centers/user-experience-center/measuring-difficulty-doesnt-need-be-difficult`, 2021. [Online; accessed 4-March-2022].

[43] J. Lehmann, M. Lalmas, E. Yom-Tov, and G. Dupret. Models of user engagement. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*, pages 164–175, UMAP'12, Springer-Verlag, Berlin, Heidelberg, 2012.

[44] J. R. Lewis, B.S. Utesch, and D.E. Maher. When there's no time for the sus. In *Proceedings of of CHI 2013*, Paris, France, 2013.

[45] J.R. Lewis. Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The asq. *SIGCHI Bulletin*, 23:78–81, 1991.

[46] J.R. Lewis. User satisfaction questionnaires for usability studies: 1991 manual of directions for the asq and pssuq (tech. report 54.609). *Boca Raton, L: International Business Machines Corporation*, 1991.

[47] J.R. Lewis. Psychometric evaluation of the computer system usability questionnaire: The csuq (tech. report 54.723). *Boca Raton, FL: International Business Machines Corporation*, 1992.

[48] J.R. Lewis. Psychometric evaluation of the post-study system usability questionnaire: The pssuq. In *Proceedings of the Human Factors Society 36th Annual Meeting*, pages 1259–1263, Santa Monica, CA: Human Factors Society, 1992.

[49] J.R. Lewis. Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7:57–78, 1995.

[50] J.R. Lewis. Umux-lite: when there's no time for the sus. In *Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 2099–2102, 2013.

[51] J.R. Lewis. Measuring perceived usability: The csuq, sus, and umux. *International Journal of Human-Computer Interaction*, 34:1148–1156, 2018.

[52] J.R. Lewis and J. Sauro. The factor structure of the system usability scale. In *Kurosu M. (eds) Human Centered Design. HCD 2009. Lecture Notes in Computer Science*, volume 5619 Lecture Notes, 2009.

[53] Y.L. Liu and C.D. Wickens. Mental workload and cognitive task automaticity - an evaluation of subjective and time-estimation metrics. *Ergonomics*, 37:1843–1854, 1994.

[54] L. Longo. Subjective usability, mental workload assessments and their impact on objective human performance. In *IFIP Conference on Human-Computer Interaction*, 2017.

[55] L.M. Lucenti. Usability Testing: cosa sono e perché si fanno i test di usabilità? `https://www.nois3.it/2021/03/`

usability-testing-cosa-sono-e-perche-si-fanno-i-test-di-usabilita/, 2021. [Online; accessed 4-March-2022].

[56] R.J. Lysaght and S.G. Hill. Operator workload: comprehensive review and evaluation of operator workload methodologies. Standard, U.S. Army Research Institute for the Behavioural and Social Sciences, 1989.

[57] Justin Mifsud. Usability Metrics - A Guide To Quantify The Usability Of Any System. https://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/, 2018. [Online; accessed 4-March-2022].

[58] S. Miller. Workload measures. *National Advanced Driving Simulator*, 2001.

[59] Jakob Nielsen. A 100-Year View of User Experience. https://www.nngroup.com/articles/100-years-ux/, 2017. [Online; accessed 4-March-2022].

[60] C. Nikulin, G. Lopez, and E. Pinonez. Nasa-tlx for predictability and measurability of instructional design models: case study in design methods. *Education Tech Research*, 67:467–493, 2019.

[61] J.M. Noyes and D.P. Bruneau. A self-analysis of the nasa-tlx workload measure. *Ergonomics*, 50(4):514–519, 2007.

[62] H.L. O'Brien and E.G. Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6).

[63] P. Park and D. Cha. *Comparison of Subjective Mental Workload Assessment Techniques for the Evaluation of In-Vehicle Navigation System Usability*. PhD thesis, Ajou University, 1998.

[64] Ammy Phuwanartnurak. Making Usability Testing work in an Agile Environment. https://medium.com/ux-uw/

making-usability-testing-work-in-an-agile-environment-d216c0338e99, 2012. [Online; accessed 4-March-2022].

[65] G.B. Reid and T.E. Nygren. The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, Advances in Psychology*, volume 52, pages 185–218, 1988.

[66] Aaron Rich and Mick McGee. In *Proceedings of the human factors and ergonomics society 48th annual meeting*, 2004.

[67] Aaron Rich and Mick McGee. Expected usability magnitude estimation. 2004.

[68] Susana Rubio, Eva DÃaz, JesÃºs MartÃn, and JosÃ© M. Puente. Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psycholgy: an International Review*, 53:61–86, 2004.

[69] B. Saket, A. Endert, and J. Stasko. Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pages 133–142, New York, NY, USA, 2016.

[70] Khalid Saleh. The 7 Step Roadmap For Effective Usability Testing. https://www.invespcro.com/blog/usability-testing/, 2017. [Online; accessed 4-March-2022].

[71] J. Sauro and J.S. Dumas. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the 27th International Conference on HumanFactors in Computing Systems*, Boston, 2009.

[72] J. Sauro and J. Lewis. *Quantifying the User Experience: Practical Statistics for User Research.* 2016.

[73] Jeff Sauro. Can You Use a Single Item to Predict SUS Scores? https://measuringu.com/single-item-sus/, 2008. [Online; accessed 4-March-2022].

[74] Jeff Sauro. Click versus Clock: Measuring Website Efficiency. https://measuringu.com/click-clock/, 2011. [Online; accessed 4-March-2022].

[75] Jeff Sauro. What Is A Good Task-Completion Rate? https://measuringu.com/task-completion/, 2011. [Online; accessed 4-March-2022].

[76] Jeff Sauro. 10 Benchmarks for User Experience Metrics. https://measuringu.com/ux-benchmarks/, 2012. [Online; accessed 4-March-2022].

[77] Jeff Sauro. Measuring Errors in the User Experience. https://measuringu.com/errors-ux/, 2012. [Online; accessed 4-March-2022].

[78] Jeff Sauro. A Brief History of Usability. https://measuringu.com/usability-history/, 2013. [Online; accessed 4-March-2022].

[79] P. Schmutz, S. Heinz, Y. M'etrailler, and K. Opwis. Cognitive load in ecommerce applications: Measurement and effects on user satisfaction. *Advances in HumanComputer Interaction*, 3:1–9, 2009.

[80] Mads Soegaard. The History Of Usability: From Simplicity To Complexity. https://www.smashingmagazine.com/2012/05/the-history-of-usability-from-simplicity-to-complexity/, 2012. [Online; accessed 4-March-2022].

[81] D. Tedesco and T. Tullis. A comparison of methods for eliciting post-task subjective ratings in usability testing. *Usability Professionals Association (UPA)*, pages 1–9, 2006.

[82] J.P. Tracy and M.J. Albers. Measuring cognitive load to test the usability of web sites. *Usability and Information Design*, pages 256–260, 2006.

[83] Pamela Tsang and Michael Vidulich. Time-sharing visual and auditory tracking tasks. research article. 1987.

[84] T.S. Tullis and J.N. Stetson. A comparison of questionnaires for assessing website usability. In *Proceedings of the 27th International Conference on HumanFactors in Computing Systems*, volume 7, 2004.

[85] M.A. Vidulich and P.S. Tsang. Assessing subjective workload assessment: A comparison of swat and the nasa-bipolar methods. In *Proceedings of the Human Factors Society 29th Annual Meeting*, volume 52, pages 71–75, 1985.

[86] M.A. Vidulich and P.S. Tsang. Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 31, pages 1057–1061, 1987.

[87] Pollione Marco Vitruvio. *De architectura*. 55.

[88] W.W. Wierwille and F.T. Eggemeier. Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, 35:263–281, 1993.

[89] wikipedia.org. Ergonomia. `https://it.wikipedia.org/wiki/Ergonomia`. [Online; accessed 4-March-2022].

[90] G.F. Wilson and F.T. Eggemeier. *Psychophysiological assessment of workload in multi-task environments. Multiple task performance*. 1991.

[91] M.S. Young and N.A. Stanton. Mental workload: theory, measurement, and application. In *Karwowski, W. (ed.) International encyclopedia of ergonomics and human factors*, volume 1, pages 818–821, 2006.

[92] F. Zijlstra and L. Van Doorn. *The Construction of a Scale to Measure Perceived Effort*. 1985.