Metodi e Strumenti per l'E-Testing

Vittorio Fuccella
Dipartimento di Matematica e Informatica
Università degli Studi di Salerno

Abstract

In questo contributo vengono descritti alcuni metodi e strumenti innovativi per l'*e-testing*. Con tale termine viene indicata la valutazione delle conoscenze degli studenti attraverso i mezzi dell'e-learning. Le proposte descritte in questo contributo sono orientate a migliorare sia la produzione che la fruizione di prove oggettive di profitto. Dal punto di vista della produzione, si propone un metodo per il miglioramento della qualità dei quesiti presentati in tali prove. Dal punto di vista della fruizione, viene presentato un metodo di analisi del comportamento degli studenti durante la fruizione dei test, finalizzato alla comprensione della strategia utilizzata per completare il test. Tali metodi sono stati sperimentati nell'ambito universitario attraverso la loro implementazione in un sistema di *e-testing* esistente, denominato *eWorkbook*.

1. Introduzione

Con il termine *e-testing*, denominato alternativamente *Computer Assisted Assessment* (*CAA*), si indica quel settore dell'e-learning finalizzato alla valutazione delle conoscenze degli studenti. Nell'*e-testing* si fa ampio uso delle prove oggettive di profitto, sia a scopi *sommativi* che a scopi *formativi*. Tali prove sono composte da quesiti *chiusi*, sia nello *stimolo* che nella *risposta* [9]. Una tipologia popolare di quesito chiuso è la *Multiple Choice*. I quesiti appartenenti a tale tipologia, pur comprendendo alcune varianti, sono composti da una *domanda* (lo stimolo), generalmente testuale o corredata da elementi multimediali come immagini o audio, e da una lista di *opzioni*, di cui una sola è la *risposta corretta*. Le risposte non corrette vengono indicate con il termine di *distrattori*.

Un test composto da quesiti chiusi offre l'indubbio vantaggio di poter essere corretto automaticamente dal calcolatore. La correzione e la valutazione sono dunque indipendenti dalla persona che corregge i test o che si limita a prendere atto dei risultati. Infine, gli studenti sono tutti nelle medesime condizioni, dal momento che a ciascuno di loro viene chiesto di effettuare lo stesso compito con lo stesso tempo a disposizione. Questa oggettività rende le prove di profitto libere da giudizi distortivi, come i giudizi espressi in base all'emotività e dunque le rende particolarmente indicate per selezioni pubbliche ed esami.

Per i motivi sopra descritti, negli ultimi anni le prove oggettive di profitto hanno suscitato un interesse crescente da parte dei ricercatori e vengono attualmente impiegate in modo massiccio nei sistemi di e-learning come parte integrante dei processi di apprendimento. Tuttavia, tali prove sono spesso progettate con superficialità, ignorando le indicazioni che la *docimologia* ha prodotto attraverso anni di ricerca ed esperimenti. Inoltre, alcuni studenti sono sovente intimoriti da una valutazione basata su prove oggettive: in molti temono di non poter esprimere al meglio le proprie capacità su tali prove e di non riuscire ad utilizzare una "strategia vincente", che consenta loro di utilizzare al meglio il tempo a disposizione per ultimare la prova, massimizzando il risultato finale. Infine, altri svantaggi delle prove oggettive includono i tempi lunghi necessari alla preparazione delle prove e l'impossibilità di valutare alcune abilità, come le capacità espressive degli studenti.

Sono state elaborate e descritte nel presente contributo alcune proposte finalizzate al miglioramento della produzione e della fruizione di prove oggettive di profitto basate su quesiti di tipo *multiple choice*.

Dal punto di vista della produzione dei quesiti, viene proposto un metodo per il miglioramento della loro qualità. Con il termine *qualità* intendiamo il potenziale di un quesito (indicato anche come

item, nel seguito) di discriminare efficacemente tra studenti preparati e non, e di ottenere il livello di difficoltà desiderato dal docente. Esistono alcuni modelli statistici, basati sui risultati dei test, che aiutano i docenti a valutare queste caratteristiche degli item. In questo contributo viene proposto un sistema che utilizza le misure proposte dai modelli statistici per valutare la qualità dei quesiti e consigliare al docente l'azione opportuna da intraprendere su un item già utilizzato in un test: continuare ad utilizzarlo in virtù della sua buona qualità; scartarlo a causa delle prestazioni scadenti; migliorarlo con opportuni accorgimenti.
Dal punto di vista della fruizione del test, lo studio presentato è finalizzato ad ottenere degli strumenti adeguati per la comprensione del comportamento dello studente durante il test. A tal fine, in passato sono stati condotti diversi esperimenti utilizzando il metodo *think aloud*. Tale metodo impone agli studenti di ragionare ad alta voce durante l'esecuzione del test per registrare alcuni loro atteggiamenti, alterando in tal modo il loro comportamento e rendendo meno attendibili i risultati. Il metodo proposto nel presente contributo, invece, sfruttando le più avanzate tecnologie Web-based, registra silenziosamente il comportamento dello studente durante il test. I dati ottenuti vengono poi analizzati attraverso un grafico progettato ad hoc. Attraverso l'analisi dei dati si è in grado di comprendere molti aspetti del comportamento degli studenti e, quindi, una volta che sarà stata raccolta una mole sufficiente di dati e che saranno state effettuate le dovute analisi, si potranno dare dei suggerimenti agli studenti su come affrontare nella maniera migliore i test, in modo da limitare la loro sfiducia nei confronti delle prove oggettive di profitto.
I metodi a cui si è fatto cenno sono stati sperimentati per ottenere dei risultati preliminari sulla loro efficacia e praticabilità. Sono stati implementati all'interno di specifici moduli del sistema *eWorkbook*, uno strumento di *e-testing* sviluppato presso il Dipartimento di Matematica e Informatica dell'Università degli Studi di Salerno. Per valutare il metodo per il miglioramento della qualità dei quesiti, si è proceduto ad organizzare due sessioni di test e si è verificato l'effettivo miglioramento della qualità del test tra la prima e la seconda sessione. Anche per la valutazione del comportamento degli studenti durante il test si è proceduto ad eseguire una sessione di test con il sistema *eWorkbook*. Il comportamento degli studenti è stato registrato e analizzato e sono state tratte delle conclusioni sulle strategie utilizzate dagli studenti per completare i test.
Il resto del contributo è organizzato come segue: nel paragrafo 2 viene analizzato lo stato dell'arte sui sistemi di *e-testing*, con particolare riferimento alla qualità dei quesiti. Vengono anche descritti alcuni degli esperimenti condotti in passato per la valutazione dei comportamenti degli studenti durante i test. Nel paragrafo 3 viene descritto il sistema per il miglioramento della qualità dei quesiti. Il paragrafo 4 descrive il metodo per il tracciamento e l'analisi del comportamento dello studente durante i test. Segue, nel paragrafo 5, la descrizione degli esperimenti per la valutazione dei metodi proposti. Infine, delle brevi considerazioni finali e l'anticipazione di possibili idee per lavori futuri concludono il contributo.

2. Lavori correlati

Per la valutazione dello stato dell'arte nell'ambito dell'*e-testing*, sono stati analizzati alcuni tra i principali sistemi LMS Web-based con funzionalità di creazione ed erogazione di test on-line, tra i quali si annoverano *Moodle* [18], *Blackboard* [3], *Sakai* [21] e *Questionmark* [20]. Questi sistemi spesso implementano le funzionalità di presentazione delle statistiche sui quesiti utilizzati. Le statistiche vengono mostrate attraverso dei report, ma i sistemi non offrono suggerimenti o aiuti per i docenti finalizzati al miglioramento della qualità dei quesiti con prestazioni basse.
Il modello *Item Response Theory* (*IRT*) [10] viene utilizzato in alcuni sistemi [11] o esperimenti [5, 22] per la selezione di quesiti da sottoporre agli esaminandi sulla base delle loro abilità pregresse.
Un sistema molto vicino ai nostri obiettivi dal punto di vista della qualità dei quesiti è quello proposto da Hung et al. [12]. Gli autori propongono un sistema di *e-testing* in cui la valutazione di alcune regole serve a determinare i quesiti con le prestazioni peggiori. Il modello da cui le regole sono desunte è quello dell'*Item Analysis* [13]. Tuttavia, le regole sono insufficienti a coprire adeguatamente la gamma di difetti che i quesiti possono presentare. Inoltre, tali regole non

rispecchiano adeguatamente la knowledge-base dalla quale sono desunte. Infine, il sistema non viene in alcun modo validato attraverso degli esperimenti. Il sistema descritto nel presente contributo è migliorativo del precedente nei seguenti aspetti:

1. amplia e perfeziona le regole utilizzate per la verifica dei quesiti;
2. propone suggerimenti da fornire al docente per il miglioramento della qualità dei quesiti;
3. gestisce l'incertezza delle regole;
4. è stato validato con esperimenti mirati.

Allo scopo di analizzare il comportamento dello studente durante il test, in passato sono stati condotti diversi esperimenti utilizzando sovente il metodo *think aloud*, che comporta però una alterazione del comportamento abituale dello studente. Alcuni di essi [1, 2] hanno riguardato l'abitudine dello studente di verificare la risposta data e di cambiarla con un'altra ritenuta esatta in fase di revisione. In tali esperimenti, sono stati registrati i cambiamenti *right-to-wrong* e *wrong-to-right* ed è stata analizzata la correlazione tra il loro numero e il risultato finale del test. Un'analisi simile, in altri esperimenti [15, 19] è stata effettuata misurando la correlazione tra il tempo impiegato a completare il test e il risultato finale. Un esperimento interessante [17] focalizza l'attenzione sulle strategie utilizzate da studenti con abilità differenti (valutati con "A", "C" o "E") per completare il test. Tuttavia, tale esperimento ha riguardato solo alcuni aspetti della strategia (l'abitudine di anticipare la risposta ad un quesito, quella di saltare le risposte alle domande ritenute più difficili per valutarle successivamente, ecc.). Non si trovano in letteratura esperimenti finalizzati ad analizzare il comportamento dello studente nella sua interezza. Ciò è dovuto principalmente alla mancanza di mezzi per effettuare la registrazione. Nel presente contributo viene descritto un metodo per la registrazione che utilizza alcune delle principali tecniche di estrazione della conoscenza e di visualizzazione delle informazioni.

3. Il sistema per il miglioramento della qualità dei quesiti

Il sistema descritto nel presente contributo consente di migliorare la qualità dei test analizzando i singoli item che li compongono. La qualità degli item viene determinata al termine di una sessione di test erogati ad un numero statisticamente significativo di studenti. Gli item del test sono marcati con dei dischi semaforici. Il disco verde indica che l'item è di buona qualità e può continuare ad essere utilizzato in test futuri. Un disco rosso indica che l'item ha prestazioni scadenti e deve essere scartato. Un disco giallo indica che l'item ha prestazioni scadenti, ma può essere migliorato con l'intervento del docente e continuare ad essere utilizzato per test futuri. Nell'ultimo caso, il sistema cerca anche di capire la causa delle prestazioni negative e di dare un suggerimento al tutor su come provare a modificare l'item.

Per ottenere il risultato descritto, il sistema effettua una classificazione degli item in base ad alcuni indici desunti da modelli statistici come l'*Item Analysis* e l'analisi dei *distrattori*. Tali indici sono utilizzati in un sistema di regole che stabiliscono in quale classe l'item debba essere inserito. Per gestire l'incertezza delle regole, e per la classificazione è stata utilizzata la *fuzzy logic* [23].

3.1. Knowledge-base

Discriminazione e *difficoltà* sono gli indici più utilizzati nei modelli statistici per la valutazione della qualità degli item. Essi possono essere utilizzati sia per determinare la qualità degli item, sia per risalire alla causa delle loro prestazioni negative. La discriminazione è in genere calcolata mettendo in correlazione i risultati ottenuti sull'item con i risultati ottenuti sull'intero test; la difficoltà è la percentuale di studenti che ha risposto in modo errato al quesito. Altri indici utilizzati sono la discriminazione e la frequenza dei *distrattori* e delle astensioni, che ci consentono di capire quali studenti (più o meno preparati) e in che misura hanno scelto una determinata opzione o si sono astenuti dal rispondere. Per ulteriori informazioni sulle statistiche relative agli item si consultino [10, 13].

Come suggeriscono gli esperti, un valore ottimale per la discriminazione di un item è pari o superiore a 0.5. Un valore inferiore a 0.2 indica prestazioni scadenti. Queste possono essere dovute

ad alcuni motivi, tra cui: l'item valuta gli studenti su argomenti che non sono oggetto del programma del corso; la domanda o le opzioni sono formulate in modo ambiguo o poco comprensibile; ecc. In genere non è semplice capire la causa delle prestazioni basse degli item solo osservando la discriminazione, quindi, in mancanza di altri indicatori, è difficile fornire un suggerimento al docente e potrebbe dover essere necessario scartare l'item. Un valore negativo per la discriminazione, se associato ad un valore positivo per la discriminazione di un *distrattore*, è segno di un possibile errore nella scelta della risposta esatta (p.e. si potrebbe essere trattato di un errore in fase di data entry). In questo caso è possibile recuperare l'item cambiando la risposta esatta.
Se la difficoltà di un item è troppo alta (>0.85) o troppo bassa (<0.15), c'è il rischio di non valutare correttamente gli studenti sull'argomento desiderato. Questo si verifica in particolare quando tali valori per la difficoltà sono affiancati da valori medio/bassi per la discriminazione. Inoltre, nel sistema presentato, il docente ha la possibilità di definire la difficoltà prevista per un item. Quanto più la difficoltà prevista per l'item si avvicina a quella rilevata dal sistema, tanto più l'item è considerato "affidabile". Una difficoltà troppo elevata o sottostimata può essere dovuta alla presenza di un *distrattore* (che si nota per la sua frequenza) troppo plausibile (tende ad attirare troppi studenti, anche quelli preparati). La rimozione o la sostituzione di tale *distrattore* può aiutare ad ottenere un item migliore. Altre volte la stima della difficoltà da parte del docente è semplicemente errata ed è necessario modificarla per un uso corretto degli item nei test, come ad esempio la creazione di test correttamente bilanciati, comprendenti banchi di item facili, di media difficoltà e difficili.
Per quanto riguarda i *distrattori*, essi possono contribuire a comporre un buon item nel momento in cui sono scelti da un numero significativo di studenti. Quando la loro frequenza è troppo elevata, ci potrebbe essere un'ambiguità nella formulazione della domanda o del *distrattore* stesso. La discriminazione di un *distrattore* dovrebbe essere negativa per indicare che il *distrattore* è scelto dagli studenti meno preparati. In conclusione, un buon *distrattore* è quello selezionato da un numero non elevato ma significativo di studenti non preparati.
Un'elevata astensione (ove consentita) è sempre sintomo di difficoltà dell'item. Quando è accompagnata da un valore elevato (non negativo prossimo allo 0) per la discriminazione dell'item, può indicare che l'item ha una qualità scadente e potrebbe non essere semplice migliorarla.

3.2. Il sistema

Quella utilizzata dal sistema che descriveremo è un tipico esempio di *fuzzy classification*, approccio già impiegato in molte applicazioni tecnologiche nei settori più disparati, dalle previsioni meteo [4] alle diagnosi mediche [8]. Il sistema per la valutazione della qualità degli item è basato su regole: le regole usano, come *variabili linguistiche*, gli indici statistici descritti nel paragrafo 3.1, calcolati al termine di una "sessione di test". Con quest'ultimo termine indichiamo la somministrazione del test ad un numero statisticamente significativo di studenti, il cui valore è impostato all'interno della configurazione del sistema.
Il sistema funziona effettuando una classificazione degli item. Sono state individuate una serie di classi a cui gli item possono essere assegnati e a ciascuna classe è stata associata una regola di produzione. Il *grado di verità* di una regola indica la misura in cui l'item appartiene alla classe associata. La classificazione è effettuata selezionando la classe con il *grado di verità* maggiore.
La knowledge-base descritta nel paragrafo precedente è stata tradotta nel set di dieci regole, mostrate in tabella 1. Le prime tre colonne della tabella contengono, rispettivamente, la classe dell'item, la regola usata per la classificazione e lo stato dell'item. Per le classi il cui stato è giallo, la quarta colonna contiene una descrizione del problema che riguarda l'item e il suggerimento per risolverlo, al fine di migliorarne la qualità.
Le regole utilizzano nove variabili linguistiche. Queste non saranno descritte nel dettaglio per brevità. Ogni variabile ha due o tre termini associati (ad esempio, alla variabile *difficulty* sono stati associati i termini *medium*, *very_low* e *very_high*), ciascuno dei quali è un insieme *fuzzy*. Per le

funzioni di appartenenza sono state utilizzate forme triangolari e trapezoidali. La maggior parte dei valori delle basi e dei vertici sono stati desunti dalla knowledge-base. Solo alcuni di loro sono stati determinati su base sperimentale. Al termine di una sessione di test, ogni item del test è classificato in una delle dieci classi descritte nella tabella e il sistema segnala al docente lo stato corrispondente. Il docente dovrà prendere la decisione opportuna, mantenere, modificare o eliminare l'item, e potrà continuare ad utilizzare il sistema, ottenendo così una migliore qualità dei quesiti utilizzati.

Tabella 1. Classificazione degli item

Classe	Regola	Stato	Problema e Suggerimento
1	`discrimination IS high AND abst_discrimination IS negative WITH 0.9`	Verde	/
2	`discrimination IS low AND abst_frequency IS high AND abst_discrimination IS positive`	Rosso	/
3	`difficulty IS very_low AND discrimination IS low`	Rosso	/
4	`difficulty IS very_high AND discrimination IS low AND max_distr_freq IS high`	Giallo	Item troppo difficile a causa di un distrattore troppo plausibile, eliminare o sostituire il distrattore *x*.
5	`difficulty_gap IS overestimated AND discrimination IS low`	Giallo	Difficoltà dell'item sovrastimata, evitare distrattori troppo plausibili e risposte troppo ovvie.
6	`difficulty_gap IS overestimated AND discrimination IS NOT low`	Giallo	Difficoltà dell'item sovrastimata, modificare la stima della difficoltà.
7	`difficulty_gap IS underestimated AND max_distr_freq IS high`	Giallo	Difficoltà dell'item sottostimata a causa di un distrattore troppo plausibile, eliminare o sostituire il distrattore *x*.
8	`difficulty_gap IS underestimated AND max_distr_freq IS NOT high`	Giallo	Difficoltà dell'item sottostimata, modificare la stima della difficoltà.
9	`max_distr_discr IS positive AND discrimination IS negative`	Giallo	Risposta corretta non indicata correttamente, selezionare l'opzione *x* come risposta corretta.
10	`discrimination IS high AND max_distr_discr IS positive AND distr_freq IS NOT low`	Giallo	Distrattore troppo plausibile, eliminare o sostituire il distrattore *x*.

4. Analisi del comportamento dello studente durante il test

La tecnica descritta in questo paragrafo consiste nel registrare le interazioni degli studenti con il sistema di *e-testing*. Più precisamente, vengono catturate le occorrenze di eventi di browsing e i responsi dati ai quesiti durante la fruizione del test. I dati raccolti sono utilizzati per la visualizzazione di un grafico che rappresenta un'analisi cronologica del test. Tale analisi è in grado di comunicare all'osservatore tutte le informazioni salienti sulla fruizione del test.

I dati acquisiti attraverso il sistema hanno una certa affidabilità, dal momento che esso ci consente di registrare il comportamento degli studenti senza che questi siano informati della registrazione, quindi si evitano comportamenti artefatti dovuti alla necessità di esprimere pensieri a voce alta, come avviene nel metodo di registrazione *think aloud*. Il sistema utilizza le più recenti tecnologie per il Web, come *AJAX*, acronimo di *Asynchronous JavaScript and XML*, che consente alle applicazioni Web-based di ottenere una maggiore interattività.

Una componente fondamentale del sistema è il *Framework di Logging* che, una volta catturate le interazioni lato client, le invia al server per la registrazione. Un'altra componente fondamentale del sistema è l'applicazione per la produzione dei grafici. Allo scopo di dimostrare l'efficacia del sistema, questo è stato sperimentato nell'ambito di un corso universitario per somministrare un test agli studenti, valido come esame finale del corso.

4.1. Il Framework di logging

L'obiettivo del *Framework di Logging* è di registrare tutte le azioni dello studente durante la fruizione del test e di memorizzare i dati in un file in formato *XML*. Il framework è composto da due componenti, una lato client ed una lato server. La prima è responsabile della cattura degli eventi che avvengono nel browser e della loro trasmissione al modulo lato server, la seconda della gestione dei file di log, in particolare della loro creazione e della registrazione degli eventi al loro interno.
Il lavoro del modulo lato client è eseguibile in qualunque Web browser, senza la necessità di dover installare alcun plug-in. Gli eventi catturati sono i seguenti:

- Azioni intraprese sulla finestra browser (open, close, resize, load, unload);
- Azioni intraprese all'interno dell'area di lavoro del browser (pressione di tasti, scrolling, movimenti e click del mouse).

Gli eventi vengono inviati al modulo lato server ad intervalli regolari. Al termine di ogni sessione utente, il documento XML viene scritto sul disco. Il modello di informazione utilizzato per il log è molto semplice: le informazioni sono organizzate per sessione dell'utente. A tale livello, vengono memorizzati lo username, l'indirizzo IP della macchina su cui il test viene eseguito e l'identificatore della sessione del browser, oltre alle informazioni sul browser (tipo, versione e sistema operativo su cui gira). All'interno di una sessione viene memorizzata la lista degli eventi. I dati relativi agli eventi sono i seguenti:

- Tipo di evento;
- Oggetto HTML sul quale è stato registrato l'evento (se presente);
- Informazioni riguardanti il mouse (bottone premuto, coordinate del puntatore)
- Informazioni riguardanti il tempo (timestamp dell'evento)
- Ulteriori informazioni specifiche di ciascun evento. Per esempio, per un evento di tipo responso (risposta ad un quesito), vengono memorizzati gli identificatori del quesito e dell'opzione selezionata e l'indicazione della correttezza della risposta data.

4.2. Visualizzazione del test

Il grafico mostra un'analisi cronologica del test, ottenuta mostrando i punti salienti della fruizione del test, sintetizzata nella registrazione delle interazioni. Il grafico mostra, istante per istante, l'item visualizzato dallo studente, la posizione del mouse (intesa come la presenza del puntatore sull'area della domanda o sulle aree delle opzioni), la presenza di interazioni di tipo responso, corretto o errato.
Il grafico è bidimensionale. L'asse orizzontale riporta una misura continua, il tempo, mente l'asse verticale mostra delle categorie, e cioè il numero progressivo dell'item correntemente visualizzato dallo studente.
L'esecuzione del test è rappresentata da una linea spezzata. L'analisi di un item (lettura della domanda e delle possibili opzioni e l'eventuale tempo necessario a rispondere) per un certo intervallo di tempo da parte dello studente è mostrata attraverso un segmento tracciato dal punto corrispondente all'inizio della visualizzazione a quello corrispondente alla sua fine. Di conseguenza, la lunghezza del segmento è proporzionale alla durata della visualizzazione dell'item corrispondente. Un segmento verticale rappresenta un evento di browsing (item precedente, item successivo).
I responsi dati da uno studente ad un item sono rappresentati attraverso dei cerchi. Il numero dell'opzione scelta è stampato all'interno del cerchio. L'indicazione della correttezza del responso è data dal colore del cerchio, che è blu per un responso corretto e rosso per un responso errato.
L'analisi grafica di un test di esempio è mostrata in figura 1. Il test è composto di 25 item e la durata massima prevista è di 20 minuti, anche se lo studente ha completato ed inviato il test in circa 17 minuti. Analizzato nella sua interezza, il grafico mostra la strategia adottata dallo studente per effettuare il test: l'esecuzione del test è evidentemente divisa in due *fasi* distinte. Nella prima, della durata di circa 9 minuti, lo studente ha analizzato tutti gli item dall'1 al 25. In questa fase sono stati

forniti responsi a 19 quesiti. Alcuni item presentano più di un responso, dovuti ad un cambiamento di idea da parte dello studente, mentre alcuni item sono stati lasciati senza risposta.

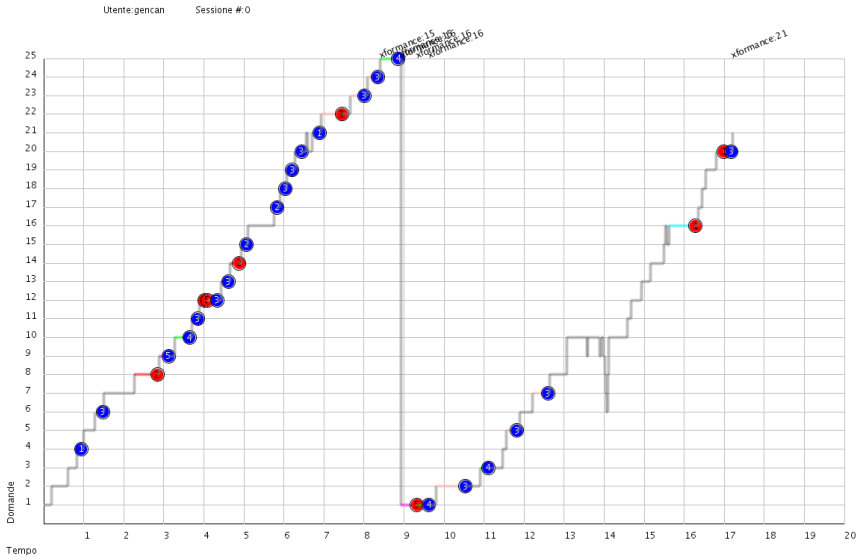

Figura 1. Analisi grafica dell'esecuzione di un test

5. Esperimenti

Gli esperimenti presentati in questo capitolo sono due. Nel primo si valuta il miglioramento della qualità dei quesiti ottenuta attraverso l'uso del metodo descritto nel paragrafo 3. Nel secondo si utilizza il metodo descritto nel paragrafo 4 per determinare le strategie usate dagli studenti per completare i test on-line. Entrambi i metodi, per poter essere sperimentati, sono stati prima implementati in appositi moduli del sistema *eWorkbook*, descritto di seguito.

5.1. Il sistema eWorkbook

eWorkbook è un sistema di *e-testing* basato sul Web, utilizzato per valutare le conoscenze degli studenti attraverso la creazione (da parte dei docenti) e la fruizione (da parte degli studenti) di test on-line. I quesiti sono memorizzati in un *repository* gerarchico. I test sono composti di una o più sezioni di quesiti. Esistono due tipologie di tali sezioni: *statiche* e *dinamiche*. La differenza consiste nel modo in cui esse permettono di selezionare i quesiti per comporre i test: per una sezione *statica*, i quesiti sono scelti direttamente dal docente durante la composizione del test. Per una sezione *dinamica*, al momento della composizione del test il docente definisce solo alcuni parametri di selezione, come la difficoltà dell'item, lasciando che il sistema li scelga in modo casuale tra quelli che soddisfano i parametri specificati, ogni volta che uno studente accede al test. In tal modo, con *eWorkbook* è possibile creare test con sezioni di item di difficoltà differente, ottenendo una difficoltà del test bilanciata, che consente di valutare al meglio una classe eterogenea dal punto di vista della preparazione degli studenti.
Il sistema è completamente accessibile con i più comuni browser Web. Non sono necessarie installazioni di plug-in o moduli aggiuntivi, dal momento che le pagine sono in HTML standard.

5.2. Miglioramento della qualità degli item

Il Sistema *eWorkbook* è stato esteso con l'aggiunta di un modulo che implementa il sistema per il miglioramento della qualità degli item descritto nel paragrafo 3. Si è poi proceduto ad una sperimentazione per valutare tale miglioramento. Un primo esperimento è consistito nell'utilizzo del sistema attraverso due sessioni di test nell'ambito di un corso universitario e misurando il miglioramento della qualità in termini di capacità di discriminazione e di rispondenza della

difficoltà a quella stimata dal docente. Un set di 50 item è stato predisposto per l'esperimento. Nella prima sessione, un test contenente 25 item scelti a caso dal set è stato somministrato a 60 studenti. Successivamente, lo stato degli item è stato verificato attraverso l'interfaccia del sistema per identificare quali modificare e quali sostituire. Una volta effettuate modifiche e sostituzioni, il test è stato somministrato ad altri 60 studenti.
In Figura 2a si mostra la tabella, esportata in un foglio di calcolo, contenente il report degli item presentati nella prima sessione e le loro prestazioni. Quelli da eliminare sono marcati in rosso, mentre quelli da modificare sono marcati in giallo: 5 item su 25 sono da scartare e 4 da modificare. Non si entrerà nel dettaglio dei problemi riscontrati e delle modifiche effettuate per ciascun quesito. Si rimanda per tali ed altri approfondimenti all'articolo [6].
È stato preparato un nuovo test, contenente gli stessi item del precedente, eccetto per i 5 scartati, sostituiti da 5 nuovi item scelti in modo casuale dai rimanenti contenuti nel set, e per i 4 modificati, sostituiti con la nuova versione ottenuta dalla modifica della precedente secondo i suggerimenti del sistema. Al termine della seconda sessione, la nostra attenzione è stata focalizzata sulla valutazione del miglioramento ottenuto in termini di qualità degli item, piuttosto che nella ricerca di nuovi item da scartare o modificare.
La Figura 2b mostra il report della seconda sessione di test. I valori della discriminazione e della difficoltà, cambiati rispetto alle stesse righe della tabella relativa alla prima sessione, sono marcati in giallo.
Per misurare il miglioramento qualitativo del test dalla prima alla seconda sessione, sono stati calcolati i seguenti parametri per ciascuno dei due test:

1. La media dell'indice di discriminazione degli item del test;
2. La media della differenza tra la difficoltà stimata e quella calcolata per gli item del test.

Per quanto riguarda il parametro 1, abbiamo osservato un miglioramento da un valore di 0.375, ottenuto nella prima sessione, ad un valore di 0.459, ottenuto nella seconda. L'incremento percentuale è del 22.4%. Per quanto riguarda il parametro 2, abbiamo registrato un decremento nella differenza media tra la difficoltà stimata dal tutor e quella calcolata dal sistema del 17.8%, passando da un valore di 0.19 a 0.156 tra le due sessioni.

Question Id	Options	Correct	Discrimination	Tutor Diff	Difficulty
1-F-1	5	1	-0,04	0,7	0,76
1-B-10	5	3	0,51	0,5	0,24
1-B-6	5	4	0,6	0,5	0,58
1-A-19	5	4	0,22	0,5	0,29
1-D-2	5	2	0,7	0,5	0,82
1-B-4	5	5	0,42	0,5	0,34
1-A-13	5	1	0,33	0,3	0,08
1-F-4	5	3	0,32	0,5	0,79
1-B-18	5	4	0,55	0,3	0,53
1-A-15	5	5	0,37	0,5	0,66
1-B-2	5	5	0,59	0,5	0,45
1-E-4	5	4	0,4	0,5	0,79
1-C-1	5	1	0,07	0,5	0,39
1-B-12	5	4	0,48	0,3	0,74
1-A-24	5	2	0,36	0,3	0,37
1-D-3	5	5	0,15	0,5	0,53
1-C-4	5	1	0,16	0,5	0,74
1-B-16	5	3	0,41	0,5	0,76
1-A-9	5	3	0,57	0,3	0,53
1-B-20	5	5	0,38	0,3	0,47
1-A-2	5	2	0,21	0,3	0,34
1-B-8	5	5	0,49	0,5	0,29
1-C-5	5	1	0,17	0,7	0,87
1-B-15	5	4	0,52	0,5	0,42
1-E-1	5	4	0,44	0,5	0,87
Average Discrimination			0,3752		
Average Difficulty Gap			0,19		

(a)

Question Id	Options	Correct	Discrimination	Tutor Diff	Difficulty
1-F-1	5	1	0,48	0,7	0,67
1-B-10	5	3	0,51	0,5	0,24
1-B-6	5	4	0,6	0,5	0,58
1-D-4	5	3	0,08	0,5	0,76
1-D-2	5	2	0,7	0,5	0,82
1-B-4	5	5	0,42	0,5	0,34
1-A-13	5	1	0,33	0,3	0,08
1-F-4	5	3	0,47	0,5	0,43
1-B-18	5	4	0,55	0,3	0,53
1-A-15	5	5	0,37	0,5	0,66
1-B-2	5	5	0,59	0,5	0,45
1-E-4	5	4	0,4	0,5	0,79
1-A-17	5	5	0,44	0,3	0,45
1-B-12	5	4	0,48	0,3	0,74
1-A-24	5	2	0,36	0,3	0,37
1-D-3	5	5	0,15	0,5	0,53
1-A-23	5	3	0,38	0,3	0,32
1-B-16	5	3	0,41	0,7	0,76
1-A-9	5	3	0,57	0,3	0,53
1-B-20	5	5	0,38	0,3	0,47
1-F-3	5	5	0,76	0,7	0,72
1-B-8	5	5	0,49	0,5	0,29
1-B-5	5	5	0,66	0,3	0,5
1-B-15	5	4	0,52	0,5	0,42
1-E-1	5	4	0,37	0,5	0,58
Average Discrimination			0,4588		
Average Difficulty Gap			0,1556		

(b)

Figura 2. Report delle due sessioni di test

5.3. Comprensione delle strategie degli studenti nell'esecuzione dei test

Un secondo esperimento, finalizzato alla comprensione delle strategie che usano gli studenti per effettuare test on-line, è consistito nella somministrazione di un test in laboratorio, contenente 25

item, da completare in un tempo massimo di 20 minuti, a 71 studenti. La strategia di valutazione adottata non prevedeva alcuna penalità per risposte errate e gli studenti ne erano al corrente. Da tale prova è stato ottenuto un log di circa 4 Mb di dimensione. I grafici sono stati ottenuti e visualizzati. Dalla loro analisi è risultato che, tranne rare eccezioni, gli studenti hanno utilizzato le seguenti tre strategie per completare il test:

- **Singol Phase**. Questa strategia è composta di una sola fase. Il tempo disponibile per completare il test è gestito dallo studente in modo da visualizzare tutti i quesiti sequenzialmente una sola volta. Lo studente tenta di ragionare su un quesito per un tempo adeguato a produrre un responso nella maggioranza dei casi. Possono essere presenti fasi successive alla prima di durata trascurabile e prive di responsi. Un esempio della strategia *Single Phase* è mostrato in figura 3a.
- **Active Revising**. Questa strategia è composta da due o più fasi. Lo studente visualizza intenzionalmente tutti i quesiti il più velocemente possible. Il tempo rimanente è utilizzato per una o più fasi di revisione. Inizialmente, lo studente non fornisce responso ai quesiti sui quali ha incertezza, lasciandolo a fasi successive. Come regola generale, la prima fase dura più tempo delle successive, che hanno durate decrescenti. Un esempio della strategia *Active Revising* è mostrato in figura 3b.
- **Passive Revising**. Questa strategia è composta da due o più fasi. Lo studente visualizza e risponde a tutti i quesiti nel più breve tempo possibile. Il tempo rimanente è utilizzato per una o più fasi di revisione. Come regola generale, la prima fase dura più tempo delle successive, che hanno durate decrescenti. Un esempio della strategia *Active Revising* è mostrato in figura 3c.

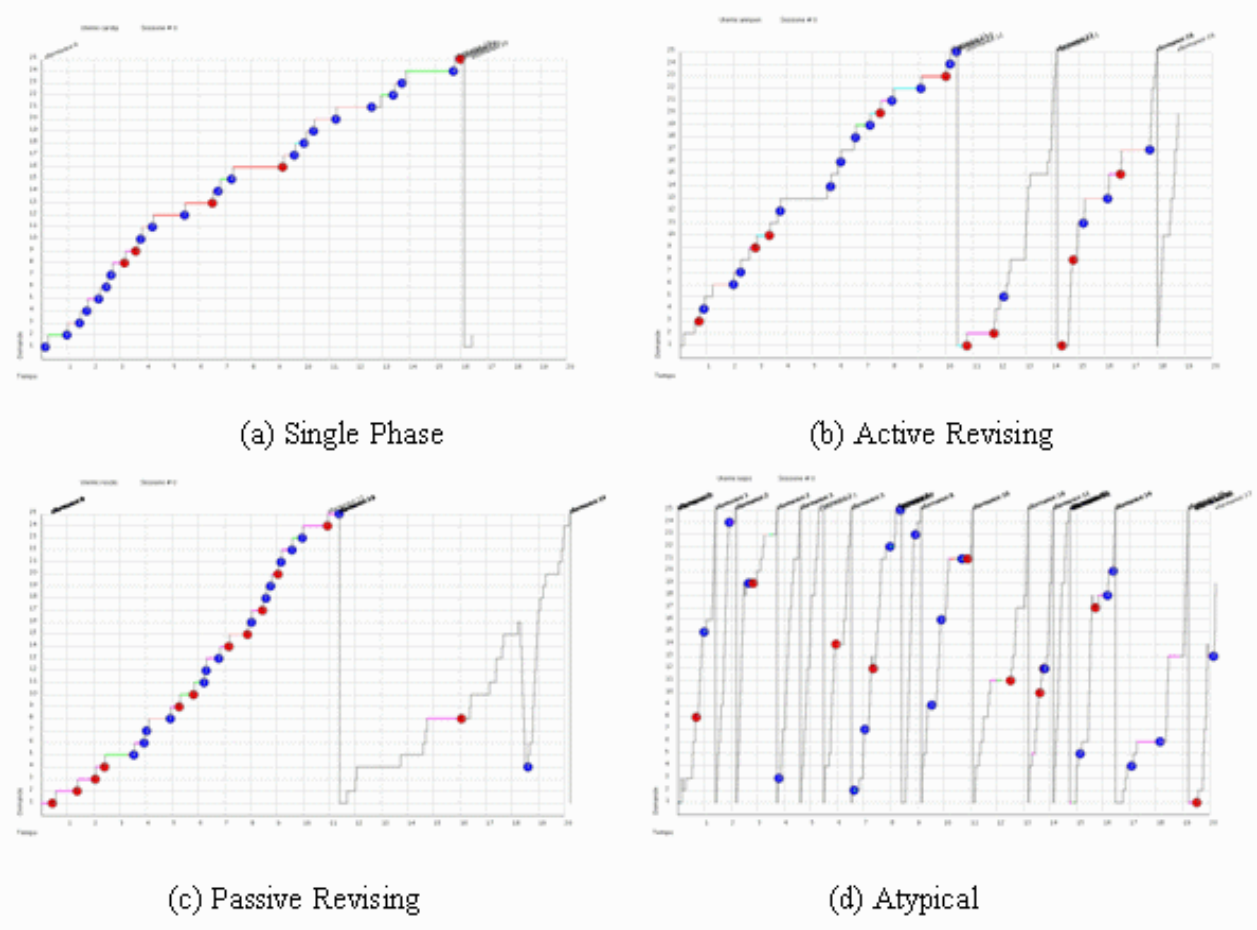

Figura 3. Esempi di strategie utilizzate per l'esecuzione dei test

Sia per la definizione delle strategie che per la classificazione delle istanze dei test in ciascuna di esse, i grafici sono stati analizzati da un operatore umano. Tali task sono abbastanza difficilmente eseguibili automaticamente, mentre un operatore umano con una certa abitudine può stabilire la strategia utilizzata attraverso la semplice osservazione dei grafici delle istanze dei test e, dopo un'analisi approfondita delle strategie, fornire suggerimenti agli studenti per avere performance migliori sui test con quesiti *multiple choice*.
Secondo quanto osservato nell'esperimento, la strategia più adottata è la *Active Revising*, utilizzata da 40 studenti su 71 (56.5%), seguita dalla *Passive Revising* (20 studenti su 71, pari al 28.2%) e dalla *Single Phase*, utilizzata solo in 9 casi su 71 (12.7%). Soltanto due studenti hanno utilizzato una strategia atipica, che non può essere ricondotta a nessuna di quelle precedentemente descritte.
I risultati migliori sono stati ottenuti dagli studenti che hanno adottato la *Passive Revising*, con una media di 17.6 responsi esatti ai 25 item del test. Con la *Active Revising*, invece, è stato ottenuto un punteggio medio di 16.4. Infine, la strategia *Single Phase* si è rivelata la peggiore, con un punteggio medio di 15.1. Pertanto, una strategia vincente prevede l'utilizzo di più di una fase. Ciò è anche confermato dalla correlazione lineare positiva (0.14) tra il numero di fasi e il punteggio ottenuto sul test. Inoltre, per entrambe le strategie che prevedono più di una fase, spesso il punteggio viene migliorato attraverso l'esecuzione di una nuova fase. Tale miglioramento è significativo nelle prime fasi e tende a diventare trascurabile al crescere del numero di fase: partendo da un punteggio medio di 14.3 ottenuto al termine della prima fase, tale valore incrementa a 16.2 al termine della seconda fase. La presenza di ulteriori fasi porta il punteggio medio a 16.5. La durata media della prima fase della strategia *Passive Revising* (14'50") è superiore a quella registrata per la strategia *Active Revising* (11'51"). Tale risultato era prevedibile, dal momento che, per definizione, la strategia *Active* prevede lo skip (ragionamento più breve) dei quesiti sui quali si ha incertezza. Un altro risultato prevedibile, dovuto ai precedenti ragionamenti, è che la strategia *Passive* ha meno fasi della *Active*, in media (2.55 e 3.2, rispettivamente).
Per ulteriori dettagli sull'esperimento si consiglia la lettura di [7].

6. Conclusioni

In questo contributo sono stati presentati alcuni metodi e strumenti per il miglioramento della valutazione on-line. I miglioramenti sono stati ottenuti sia dal punto di vista della qualità che dal punto di vista della fruizione dei test on-line: sotto il primo aspetto si è riusciti ad ottenere dei test formati da item qualitativamente migliori; per la fruizione, è stato analizzato il comportamento dello studente durante il test e sono state desunte e analizzate le strategie utilizzate dagli studenti per completare il test. In entrambi i casi sono state eseguite delle sperimentazioni: per valutare l'effettivo miglioramento qualitativo, nel primo caso, e per determinare e classificare le strategie utilizzate, nel secondo caso. I risultati delle sperimentazioni sono promettenti e incoraggiano a proseguire il lavoro svolto.
Il sistema proposto utilizza un metodo di *fuzzy classification* per la classificazione dei quesiti. In futuro, si prevede l'impiego di altri metodi, come il metodo delle k-medie [16] o alcuni metodi gerarchici [14] e di correlazione [4] e di effettuare una valutazione comparativa tra le loro prestazioni nel caso specifico. Essendo i metodi citati basati sui dati, per poter essere utilizzati necessitano della disponibilità di un'ampia base di informazioni per i risultati dei test. Tali informazioni sono al momento in fase di raccolta con l'utilizzo del sistema *eWorkbook* in alcuni corsi universitari.
Per l'analisi dei test attraverso la visualizzazione, si sta cercando di inserire i grafici all'interno di un'interfaccia interattiva che ne faciliti la lettura. Inoltre, si sta pensando di utilizzare gli stessi grafici per altre applicazioni, come la scoperta di eventuali correlazioni tra i quesiti di uno stesso test e di comportamenti truffaldini nello svolgimento di test da parte degli studenti.

7. Riferimenti

[1]. Bath, J.A. (1967). *Answer-changing Behaviour on objective examinations*. The Journal of Educational Research. 61. pp 105-107.

[2]. Best, J.B. (1979). *Item difficulty and answer changing.* Teaching of Psychology, 6, pp. 228-240.

[3]. *Blackboard* (2007). http://www.blackboard.com/

[4]. Bradley R.S., Barry R.G., Kiladis G. (1982). *Climatic fluctuations of the western United States during the period of instrumental records*. Final report to the National Science Foundation, University of Massachussets, Amherst.

[5]. Chen, C.M., Duh, L.J., Liu, C.Y. (2004). *A Personalized Courseware Recommendation System Based on Fuzzy Item Response Theory*. In Proceedings of IEEE Int. Conf. on e-Technology, e-Commerce and e-Service. pp. 305-308

[6]. Costagliola G., Ferrucci F., Fuccella V. (2007) *A Web-Based E-Testing System Supporting Test Quality Improvement*. Proceedings of The 6th International Conference on Web-based Learning. Pp 190 - 197

[7]. Costagliola G., Fuccella V., Giordano M., Polese G. (2007), *Logging and Visualizing Learner Behavior in Web-Based E-testing*. Proceedings of The 6th International Conference on Web-based Learning. Pp 272 – 279.

[8]. Exarchos T. P., Tsipouras M. G., Exarchos C. P., Papaloukas C., Fotiadis D. I., Michalis L. K. (2007). *A methodology for the automated creation of fuzzy expert systems for ischaemic and arrhythmic beat classification based on a set of rules obtained by a decision tree*. Arificial Intelligence in Medicine, 40 (3), pp. 187-200

[9]. Frignani, P., Bonazza, V. (2003). *Le Prove Oggettive di Profitto. Strumenti Docimologici per l'Insegnante*. Carocci.

[10]. Hambleton R. K., Swaminathan, H. (1985). *Item Response Theory-- Principles und Applications*, Netherlands: Kluwer Academic Publishers Group.

[11]. Ho, R.G., Yen, Y.C. (2005). Design and Evaluation of an XML-Based Platform-Independent Computerized Adaptive Testing System. *IEEE Transactions on Education*, 48 (2). pp. 230-237

[12]. Hung, J.C., Lin, L.J., Chang, W.C., Shih, T.K., Hsu, H.H., Chang, H.B., Chang, H.P., Huang, K.H. (2004). *A Cognition Assessment Authoring System for E-Learning*. In Proc. of 24th Int. Conf. on Distributed Computing Systems Workshops. pp. 262-267

[13]. *Item Analysis* (2007). http://www.washington.edu/oea/pdfs/resources/item_analysis.pdf

[14]. Johnson, S.C. (1967). *Hierarchical Clustering Schemes*. Psychometrika, 32. pp 261-274

[15]. Johnston, J.J. (1977). *Exam Taking speed and grades*. *Teaching of Psychology*, 4, 148-149

[16]. MacQueen J. (1967). *Some Methods for Classification and Analysis of Multivariate Observations*. Fifth Berkeley Symposium on Mathematics, 1. pp 281-298

[17]. McClain, L. (1983). *Behavior during examinations: A comparison of "A", "C" and "F" students.* Teaching of Psychology 10 (2).

[18]. *Moodle* (2007). http://moodle.org/

[19]. Paul, C.A.; Rosenkoetter, J.S. "The relationship between the time taken to complete an examination and the test score received.", *Teaching of Psychology*, 1980, 7, 108-109.

[20]. *Questionmark* (2007). http://www.questionmark.com/

[21]. *Sakai* (2007). http://sakaiproject.org

[22]. Sun K. T. (2000). An Effective Item Selection Method for Educational Measurement, *In Proceedings of International Workshop on Advanced Learning Technologies.* pp. 105-106.

[23]. Zadeh L. A. (1977). *Fuzzy sets and their applications to classification and clustering*. Academic Press.