La migrazione delle prove INVALSI di Matematica da PPT a CBT. Uno studio sulle prove di pre-test per la II superiore

Emanuela Botta • Università degli Studi di Roma La Sapienza – emanuela.botta@uniroma1.it
Cristina Lasorsa • INVALSI – cristina.lasorsa@invalsi.it

The migration of INVALSI mathematics tests from PPT to CBT. A study about field-trial tests for the second class of Secondary school

Questo articolo dà conto dei risultati di una sperimentazione INVALSI per il passaggio da prove cartacee a prove computer based condotta nella campagna di pre-test del 2016 di matematica del livello 10. Scopo dello studio è il confronto fra la somministrazione cartacea e quella computer based della stessa prova, sia in relazione alla stima dell'abilità degli studenti sia in relazione al comportamento degli item in termini di parametri di difficoltà, di indice di discriminatività e di percentuale di risposte omesse. A un campione rappresentativo a livello nazionale degli studenti di livello 10 è stata somministrata una prova di matematica in modalità computer based e a un campione parallelo la stessa prova in modalità cartacea. I risultati mostrano che la prova in formato computer based risulta mediamente più difficile della prova in formato cartaceo. Non ci sono invece differenze significative nella stima dell'abilità degli studenti.

Parole chiave: valutazione; abilità di un soggetto; discriminatività di un item; CBT; PPT; prove di matematica

This article describes an INVALSI trial for passing from paper based tests to computer-based tests conducted in the 2016 math pre-test campaign for the second class of the secondary school (level 10). The purpose of the study is to compare paper & pencil and computer-based delivery of the same test, in relation to the student's ability and to the item functioning in terms of difficulty (b), discrimination (Id), and percentage of missing. A nationally representative sample of 10th grade students was administered computer-based mathematics test. In addition a parallel sample of students was administered the same test in paper & pencil mode. The results show that test in computer-based mode was significantly harder statistically than the paper-based test. There are no significant differences in the estimated student's ability.

Keywords: assessment; student's ability; item discrimination; CBT; PPT; math test

* **Si ringraziano l'INVALSI per le basi dati fornite, la Prof.ssa Angela Martini e il Prof. Pietro Lucisano per i loro preziosi suggerimenti.**

La migrazione delle prove INVALSI di Matematica da PPT a CBT. Uno studio sulle prove di pre-test per la II superiore

1. Introduzione

Con la dicitura CBT (*Computer Based Test*) si intende qualunque tipo di prova di valutazione somministrata in formato digitale, sia on-line sia in locale, che *consenta di fare inferenze sulle conoscenze individuali, le attitudini, le abilità o altri costrutti sulla base delle informazioni ricavabili dai risultati* (Pepper, 2012; JISC, 2006). Nonostante i numerosi vantaggi di una prova CBT, l'implementazione di un sistema di valutazione basato su questo tipo di prove richiede che si tengano in considerazione numerosi aspetti, come le finalità della valutazione e l'identificazione dell'oggetto della misurazione, ma, in particolare, la comparabilità tra prove cartacee e prove computerizzate.

Negli ultimi anni nelle rilevazioni degli apprendimenti, sia nazionali sia internazionali, è aumentata in modo significativo la diffusione della modalità *computer based*, ma gli studi sull'efficacia e l'attendibilità delle misurazioni effettuate non danno risultati concordi.

Secondo l'OCSE, ad esempio, gli item *computer based* risultano più facili da comprendere e consentono un più ampio spettro di tipologie di risposta (PISA, 2015a e 2015b); altri sostengono invece che l'uso esclusivo di questa metodologia potrebbe inibire la possibilità di utilizzare strategie di soluzione alternative dei quesiti (Bennett et al. 2008). Emerge inoltre l'importanza di assicurarsi che gli item stiano effettivamente valutando il costrutto d'interesse e che le interferenze da altre fonti di variabilità, come la familiarità con l'uso del computer, siano irrilevanti o poco significative.

Alcune ricerche suggeriscono che il formato *computer based* può influenzare il rendimento degli studenti nelle prove di matematica, sia in positivo, perché essi trovano la risoluzione di problemi più stimolante e motivante – e ciò nonostante la poca familiarità con il tipo di problema e la natura dell'item (Richardson et al., 2002) – sia in negativo (Bennett, 2003; Johnson, Green, 2004 e 2006; Bridgeman, Lennon, Jackenthal, 2003), a causa di fattori strettamente legati al formato della domanda, alle diverse funzionalità rese disponibili su computer, alle azioni richieste per rispondere o ai problemi tecnici che si possono presentare in fase di somministrazione (Bennet, 2008; Johnson, Green, 2004, 2006).

Per contro altri studi (Wang et al., 2007; Karkee, Kim, Fatica 2010) suggeriscono che la modalità di somministrazione non ha effetti significativi sul punteggio complessivo e gli studi fatti per il PIAAC[1] nel 2012 mostrano che quasi tutti i parametri degli item si mantengono stabili al variare della modalità di somministrazione.

Un'analisi comparativa degli studi svolti fino al 2004 afferma che non è opportuno trarre conclusioni definitive sulla base di ricerche i cui risultati non possono

1 (OECD, 2012), http://www.oecd.org/skills/piaac/PIAAC%20Framework%202012-%20Revised%2028oct2013_ebook.pdf

essere banalmente generalizzati e consiglia di condurre specifici studi che tengano conto della natura delle prove e della tecnologia scelta per implementarle (Pommerich, 2004; Poggio, Glasnapp, Yang, Poggio, 2005).

In quest'ottica, nell'ambito della progettazione delle prove nazionali di rilevazione dell'INVALSI, è stata avviata la sperimentazione di cui si discutono gli esiti in questo lavoro, focalizzata sullo sviluppo di prove lineari[2] di matematica per gli studenti del secondo anno della scuola superiore e volta a minimizzare le differenze fra le due modalità di somministrazione, cartacea e *computer based*.

Le domande cui la sperimentazione ha cercato di rispondere sono essenzialmente due:

1. Gli item hanno un comportamento diverso in funzione della modalità con cui le prove vengono somministrate agli studenti?
2. Gli studenti ottengono risultati diversi in relazione alla modalità con cui hanno sostenuto la prova?

1.1. *La struttura delle prove e le condizioni di somministrazione*

Per rispondere agli obiettivi della sperimentazione sono state predisposte quattro prove parallele (T1-A, T1-B, T1-C, T1-D), ancorate fra loro da cinque item a scelta multipla. Le quattro prove sono state realizzate sia in formato cartaceo (PPT), sia in formato *computer based* (CBT), così che ognuna di esse fosse disponibile nelle due versioni, costituite entrambe dagli stessi quesiti.

Il passaggio dal formato cartaceo al formato *computer based* è stato effettuato con un approccio *migratory* (Ripley, 2009), riducendo al minimo le differenze fra le due modalità. Le domande in formato *computer based* hanno mantenuto sostanzialmente le stesse caratteristiche che avevano sulla carta, né sono state introdotte funzionalità aggiuntive che potessero permettere allo studente di interagire dinamicamente e in tempo reale con la domanda sfruttando le potenzialità della somministrazione digitale. Non è stata pertanto prevista la possibilità di operare sulle immagini disegnando o scrivendo su di esse, non sono stati introdotti video o simulazioni interattive, né applicazioni per l'elaborazione di fogli di calcolo o per la geometria dinamica. Nonostante ciò, la migrazione ha richiesto alcuni adattamenti che, per quanto possibile, sono stati estesi anche al formato cartaceo.

Ciascuna delle quattro prove era costituita da item[3] di differente formato:

– Scelta multipla semplice (*Simple Multiple Choice*: SMC)
– Aperti con risposta univoca (*Closed-Constructed Response*: CCR)
– Aperti con risposta articolata (*Constructed Response*: CR)
– Scelta multipla complessa (*Complex Multiple Choice*: CMC)
– Cloze (CL)

2 I modelli di prova standardizzata sono classificabili in base all'insieme di item proposti agli studenti in, prove lineari, in cui tutti gli studenti svolgono la medesima prova o prove equivalenti, o prove adattative, in cui ciascuno studente svolge una prova che si adegua gradualmente al suo livello di abilità. (Thompson, Weiss, 2009)

3 In questa sede si intendono convenzionalmente per item i quesiti elementari di cui si può comporre una domanda.

Le prove comprendevano sia quesiti semplici, nei quali allo stimolo faceva capo un solo item, sia quesiti a grappolo[4], nei quali ad uno stesso stimolo facevano capo più item. Nel formato *computer based* si è fatto corrispondere ogni item ad una schermata: i quesiti composti da un solo item erano visualizzabili in un'unica schermata mentre quelli con più item sono stati proposti allo studente in una serie di schermate successive, strutturate in modo che la pagina risultasse divisa in due colonne, con lo stimolo fisso nella prima colonna, a sinistra, e gli item in sequenza nella seconda colonna, a destra.

Tutti i quesiti sono stati costruiti in modo che le informazioni necessarie per rispondere fossero disponibili allo studente e non vi fosse l'esigenza di navigare all'interno della prova per recuperarle da una schermata precedente. Per gli item a risposta chiusa è stato posto il vincolo che lo studente fosse obbligato a dare un'unica risposta, impedendogli così di fornire risposte non valide, mentre per gli item a risposta aperta non sono stati posti limiti al numero di caratteri digitabili o al formato della risposta (numeri o lettere), lasciando che lo studente commettesse gli stessi errori che avrebbe potuto fare sulla carta. Inoltre, agli studenti è stato permesso di navigare all'interno della prova per la sua intera durata, consentendo loro di tornare più volte su un item e di cambiare le risposte date fino allo scadere del tempo previsto. Agli studenti sono stati forniti, in formato cartaceo, il formulario usualmente disponibile durante le prove INVALSI di matematica, le istruzioni per digitare correttamente eventuali simboli o formule matematiche e alcuni fogli bianchi per lo svolgimento dei calcoli o la riproduzione delle figure. È stato inoltre loro consentito di usare una calcolatrice.

1.2. *Il campione*

La sperimentazione è stata condotta su un campione di circa 2000 studenti rappresentativo della popolazione di alunni che frequentavano in Italia la seconda classe della scuola secondaria di secondo grado nell'anno scolastico 2015-16, selezionato con un metodo a due stadi.

Al primo stadio, in ognuna delle tre grandi aree geografiche italiane, Nord, Centro e Sud-Isole, è stato individuato un campione di giudizio di scuole (62). Al secondo stadio, in ciascuna scuola è stato selezionato un campione di classi, da due a quattro per ogni scuola.

Infine, in ogni classe, è stata effettuata un'assegnazione randomizzata degli studenti alla modalità di somministrazione della prova, su carta (PPT) o *computer based* (CBT), e a una delle quattro prove parallele (T1), tra loro – come accennato – ancorate. La tabella che segue mostra la ripartizione del campione nei vari gruppi e sottogruppi.

4 Il primo tipo di quesiti (semplici) era costituito da uno stimolo iniziale seguito da un unico item di uno dei formati sopra elencati; il secondo tipo di quesiti (a grappolo) era costituito da uno stimolo iniziale a cui erano collegati più item di vario formato, ad ognuno dei quali era attribuito un punteggio.

	CBT (N = 933)	**PPT** (N = 921)
T1-A	232	255
T1-B	232	237
T1-C	239	247
T1-D	230	182

Tab. 1: La suddivisione degli studenti campionati fra i due tipi di somministrazione e le quattro prove

Per verificare che i due gruppi di studenti, CBT e PPT, fossero effettivamente equivalenti, tutti gli studenti del campione sono stati sottoposti, tra gennaio e febbraio 2016, a una prova comune di matematica in formato cartaceo, composta da 24 item. La prova è parte di quella utilizzata dall'INVALSI per le procedure di ancoraggio fra le prove da una rilevazione alla successiva. Essa può ritenersi adeguata per una stima iniziale dell'abilità matematica degli studenti e per la definizione di una opportuna scala[5] (alpha di Cronbach = 0,76).

Sui risultati di tale prova iniziale comune (T0)[6] è stato effettuato un test *t* per campioni indipendenti sull'uguaglianza delle medie, a varianze uguali presunte. Il test è stato eseguito sui punteggi nella prova e sulle stime delle abilità degli studenti, calcolate con il modello di Rasch a 1 parametro.

Le statistiche generali (tab. 2) evidenziano che le medie e le deviazioni standard nei due gruppi di studenti sono sostanzialmente identiche e i risultati del test *t* (tab. 3) mostrano, come atteso, che non è possibile rifiutare l'ipotesi che le medie delle abilità e dei punteggi siano uguali nella popolazione, con un grado di fiducia del 95%.

	Gruppo	N	Media	Deviazione std.	Errore standard della media
Abilità	CBT	934	0,041	1,089	0,036
	PPT	951	-0,006	1,079	0,035
Punteggio	CBT	934	7,390	3,914	0,128
	PPT	951	7,210	3,740	0,121

Tabella 2 - Statistiche generali dei gruppi PPT e CBT – Prova T0

5 La prova presenta caratteristiche misuratorie stabili essendo stata utilizzata più volte, su campioni differenti di studenti, mostrando sempre un buon comportamento.

6 Come si dirà più avanti, tutte le analisi statistiche sono state effettuate solo sugli studenti del campione che hanno sostenuto sia la prova comune cartacea sia una delle quattro prove in modalità PPT o CBT e hanno risposto a più di metà delle domande. I dati riportati nelle tabelle da 2 a 11 sono dunque relativi ai soli studenti del campione originario considerati per le analisi.

Test t per l'eguaglianza delle medie							
	t	gl	Sign. (a due code)	Differenza delle medie	Errore standard della differenza	Intervallo di confidenza della differenza al 95%	
						Inferiore	Superiore
Abilità	0,95	1883	0,342	0,047	0,050	-0,051	0,145
Punteggio	1,011	1883	0,312	0,178	0,176	-0,167	0,524

Tabella 3 - Test per campioni indipendenti – Prova T0

Il test è stato effettuato anche sui sottogruppi designati a sostenere una delle quattro prove T1, fornendo risultati sostanzialmente identici tra loro, tranne che per il sottogruppo relativo alla prova T1-D, per il quale la differenza fra le medie, compresa fra 0,028 e 0,433, risulta statisticamente significativa (P-value = 0,02) anche se prossima allo zero.

2. La somministrazione delle prove nelle due modalità e i risultati

Dopo la prova comune cartacea, svolta, come si è detto, tra gennaio e febbraio del 2016, gli studenti hanno effettuato le prove nelle due modalità, PPT e CBT, tra aprile e maggio dello stesso anno. Le prove *computer based* sono state effettuate on-line tramite una piattaforma appositamente predisposta dall'INVALSI. Tutte le prove, sia nella versione cartacea sia in quella computerizzata, si sono svolte alla presenza di somministratori INVALSI, opportunamente formati.

I casi effettivamente presi in considerazione per le analisi sono quelli dei soli studenti che hanno sostenuto sia la prova iniziale comune (T0), sia una delle quattro prove (T1) nella modalità loro assegnata in base al disegno di ricerca.

Dalle analisi sono inoltre stati esclusi gli studenti che in almeno una delle prove avevano ottenuto un punteggio uguale a zero o non avevano risposto a più di metà delle domande. Gli studenti i cui dati sono stati analizzati sono 1854.

Tutti gli item delle quattro prove sono stati trattati, ai fini dell'analisi statistica, come item dicotomici.

Gli item a risposta aperta univoca o aperta articolata (richiesta di mostrare i calcoli o di fornire giustificazioni), dopo la correzione delle risposte effettuata per entrambe le modalità di somministrazione da un gruppo di correttori esperti sulla base di una rigida griglia di correzione, sono stati resi dicotomici classificando le risposte solo come "esatte" o "errate" ed escludendo la possibilità dell'attribuzione di un punteggio parziale.

Anche i quesiti a scelta multipla complessa sono stati trattati come dicotomici, assegnando loro il punteggio 1 (esatto) quando erano date 2 risposte corrette se gli item erano 3, o 3 risposte corrette se gli item erano 4. Il numero di risposte corrette da raggiungere per ciascun quesito a scelta multipla complessa è stato definito sulla base delle frequenze cumulate delle risposte fornite ai vari item.

Infine, per gli item di tipo *cloze* (in cui si chiede di riempire con il termine esatto le lacune di un breve testo), la risposta è stata ritenuta corretta solo se tutti i completamenti richiesti risultavano tali. I termini con cui riempire le lacune erano dati in un elenco di parole fra le quali scegliere ed era posto il vincolo che ciascun termine fosse utilizzato una sola volta. I completamenti risultavano quindi dipendenti tra loro.

Il gruppo di studenti cui era stata assegnata la prova in formato *computer based* ha risposto, prima di questa, a un questionario per valutarne la familiarità con l'uso del computer, costituito da 8 domande e somministrato su carta. Il gruppo di studenti cui era stata assegnata la prova su carta, invece, ha svolto solamente una delle quattro prove di matematica nel formato cartaceo.

L'analisi di ciascuna delle quattro prove T1 è stata effettuata solo sui quesiti che avevano ottenuto una percentuale di risposte corrette compresa fra il 10% e il 90% in entrambe le modalità di somministrazione. Si è ritenuto infatti che i quesiti con una percentuale di risposte corrette al di sotto o al di sopra di queste due soglie non potessero fornire né informazioni utili sull'eventuale variazione di funzionamento in relazione alla modalità di somministrazione né una misura accurata dell'abilità degli studenti. Dalle quattro prove nella forma originale, costituite ciascuna da un numero di quesiti compreso tra 25 e 27, sono stati eliminati complessivamente 31 quesiti.

Dopo le eliminazioni, la composizione delle prove sottoposte ad analisi risultava la seguente:

Prova	Item di ancoraggio	Item MC	Item RU	Item RG	Item MCC	Item cloze	Totale
T1-A	5	6	5	2	3		**21**
T1-B	5	2	8	1	2	1	**19**
T1-C	5	3	5	2	1		**16**
T1-D	5	7	4	0	2		**18**
Totale		**18**	**22**	**6**	**8**	**1**	

Tabella 4 – Composizione delle prove T1 analizzate

2.1. *La procedura di analisi degli item*

Per ognuno degli item delle prove nella modalità cartacea e *computer based* sono state calcolate la difficoltà degli item sulla scala di Rasch, la loro discriminatività (correlazione punto-biseriale) e la percentuale di mancate risposte.

Prima di procedere all'analisi degli item per rilevare la presenza di eventuali variazioni nel loro funzionamento a seconda della modalità di somministrazione della prova, si è provveduto ad ancorare esternamente le quattro prove T1 con la prova cartacea iniziale T0, al fine di ottenere stime della difficoltà degli item su una medesima scala e dunque avere dati fra loro confrontabili. L'ancoraggio è stato effettuato in due fasi:

1. in una prima fase si sono stimati i parametri degli item della prova T0 sull'intero campione;
2. in una seconda fase si è proceduto ad ancorare gli item della prova T0 con quelli delle prove T1. I parametri degli item della prova T0 sono stati vincolati ai valori stimati nella prima fase mentre i parametri degli item delle prove T1 sono stati lasciati liberi di variare per evidenziare le eventuali differenze di comportamento nelle due modalità di somministrazione.

Poiché le prove T1, essendo costituite dai medesimi item, sono a due a due equivalenti dal punto di vista del contenuto matematico (es.T1-A PPT e T1-A CBT), e poiché i vari sotto-gruppi di studenti risultano, come si è visto preceden-

temente, equivalenti in termini di abilità, le eventuali differenze riscontrate nel funzionamento degli item sono attribuibili alla diversa modalità di somministrazione. Ai fini dell'analisi si è assunto che tutte le prove, sia in versione cartacea sia in versione *computer based,* stimassero lo stesso costrutto unidimensionale.

2.2. *La difficoltà degli item*

Il comportamento degli item è stato analizzato innanzitutto in termini di parametro di difficoltà (*b*), stimato con il modello IRT a 1 parametro.

Per ogni item il parametro di difficoltà *b* è stato stimato utilizzando le risposte date dagli studenti nelle due modalità di somministrazione. L'analisi delle variazioni del valore stimato per *b* ha consentito di verificare se gli effetti dovuti alla modalità di somministrazione sono riscontrabili sugli item in modo generalizzato o se dipendono dal comportamento di alcuni item *outliers*.

La tabella che segue riassume le stime del parametro di difficoltà per tutti gli item delle prove T1 in entrambe le modalità e riporta la differenza fra il valore assunto dal parametro nella modalità *computer based* (CBT) e quello assunto nella modalità cartacea (PPT). Gli item sono posti in ordine crescente in base al valore di tale differenza.

Id	Item	Formato	CBT	PPT	CBT - PPT	Id	Item	Formato	CBT	PPT	CBT - PPT
7	**A7**	**RG**	**0,922**	**1,695**	**-0,773**	11	A11	MC	0,006	-0,047	0,053
20	**B4**	**RU**	**-1,052**	**-0,464**	**-0,588**	38	C8	RU	2,323	2,262	0,061
43	**D2**	**RU**	**1,330**	**1,858**	**-0,528**	51	D10	MC	0,863	0,773	0,090
9	**A9**	**MCC**	**0,095**	**0,523**	**-0,428**	21	B5	RU	1,476	1,350	0,126
22	**B6**	**RG**	**1,472**	**1,895**	**-0,423**	8	A8	RU	1,180	1,028	0,152
16	A16	MC	1,506	1,747	-0,241	2	A2	RU	0,035	-0,156	0,191
44	D3	MC	-0,732	-0,520	-0,212	50	D9	MC	1,535	1,330	0,205
34	C4	RU	-0,317	-0,111	-0,206	25	B9	RU	2,129	1,902	0,227
14	A14	RG	1,990	2,190	-0,200	13	A13	MCC	0,035	-0,214	0,249
12	A12	MC	0,835	1,013	-0,178	5	A5	RU	-0,344	-0,602	0,258
48	D7	MC	1,121	1,268	-0,147	15	A15	MC	0,937	0,677	0,260
4	A4	MCC	0,115	0,243	-0,128	24	B8	MCC	1,929	1,660	0,269
3	A3	MC	0,661	0,777	-0,116	41	C11	RU	2,655	2,339	0,316
45	D4	MCC	0,337	0,432	-0,095	31	C1	MC	2,370	2,027	0,343
19	B3	RU	1,045	1,118	-0,073	52	D11	MC	1,627	1,257	0,370
1	A1	MC	1,996	2,044	-0,048	46	D5	RU	1,383	0,942	0,441
32	C2	MC	0,830	0,854	-0,024	42	D1	RU	0,398	-0,124	0,522
17	B1	MC	0,445	0,469	-0,024	28	B12	CL	2,300	1,706	0,594
49	D8	MC	0,885	0,905	-0,020	27	B11	RU	3,208	2,614	0,594
18	B2	RG	1,565	1,579	-0,014	40	C10	RU	2,869	2,260	0,609
6	A6	RU	2,229	2,230	-0,001	**36**	**C6**	**RG**	**2,519**	**1,853**	**0,666**
10	A10	MCC	0,234	0,225	0,009	**35**	**C5**	**RU**	**2,573**	**1,886**	**0,687**
39	C9	MCC	-0,180	-0,214	0,034	**54**	**D13**	**MC**	**0,439**	**-0,372**	**0,811**
53	D12	MC	1,154	1,118	0,036	**47**	**D6**	**RU**	**2,010**	**1,169**	**0,841**
37	C7	MC	0,721	0,682	0,039	**29**	**B13**	**CL**	**3,394**	**2,466**	**0,928**
26	B10	RU	1,585	1,535	0,050	**30**	**B14**	**RU**	**3,357**	**2,344**	**1,013**
33	C3	RU	0,269	0,218	0,051	**23**	**B7**	**RU**	**4,724**	**2,349**	**2,375**

Tabella 5 - Stima del parametro di difficoltà *b*

Gli item che sono risultati più facili in formato *computer based* hanno per lo più in comune la caratteristica di essere gli item finali di un quesito a grappolo (vedi nota n. 3). La differenza sostanziale fra le due modalità di somministrazione per questi quesiti è la presentazione. Sulla carta infatti sono proposti in un'unica pagina, con lo stimolo all'inizio e gli item a seguire, mentre in CBT lo stimolo è ripetuto per ciascun item e questi sono visualizzati separatamente. La presentazione degli item uno alla volta sembra dunque rafforzare l'indipendenza fra un item e l'altro e facilitare la riflessione sul singolo item piuttosto che sul quesito nel suo insieme. Anche la vicinanza fisica fra lo stimolo e la singola domanda sembra influire positivamente sul risultato. Tali domande contengono infatti tabelle o grafici la cui analisi risulta indispensabile per fornire la risposta corretta.

In definitiva, sono risultati significativamente più facili in modalità CBT gli item 7, 20, 43, 9 e 22 (in grassetto nella tabella).

I primi due item (7 e 20) sono quelli che presentano la differenza più rilevante. L'item 7 è a risposta aperta articolata: allo studente viene chiesto di valutare la correttezza o meno di un'affermazione relativa al confronto fra i grafici di tre funzioni e di fornire una motivazione della scelta effettuata. La separazione di questo item dai precedenti afferenti allo stesso stimolo, in cui si richiedevano le formule delle funzioni corrispondenti ai tre grafici, ha permesso allo studente di focalizzare la propria attenzione su di essi piuttosto che sulle numerose altre informazioni presenti nel quesito, favorendo così una puntualizzazione dell'obiettivo dell'item. L'item 20 è invece a scelta multipla complessa, ma anche in questo caso tutte le affermazioni, per le quali si chiedeva di dire se fossero vere o false, sono riferite a una tabella che, nella presentazione CBT, rimane sempre presente a fianco del testo.

L'item 9 è a risposta aperta univoca. È probabile che la maggiore facilità di questo item in CBT sia dovuta alla considerevole diminuzione del numero di *missing* (17,67% di *missing* in CBT contro il 32,55% in PPT). Tale item è il secondo di due riferiti a uno stesso stimolo in cui si pone un problema di calcolo delle probabilità. Nel primo item si chiedeva una rappresentazione dell'insieme universo per elencazione degli esiti, poco utile per rispondere correttamente all'item 9. Per rispondere a questo item era infatti necessaria una diversa rappresentazione dell'insieme universo. La distanza fra i due item nella presentazione CBT ha facilitato allo studente l'individuazione della strategia di risoluzione più adatta.

Gli item 36, 35, 54, 47, 20, 30 e 23 sono invece risultati significativamente più difficili in CBT piuttosto che in modalità PPT. Ad eccezione dell'item 54, che è un quesito a scelta multipla complessa, tutti gli altri sono item a risposta aperta univoca o a risposta aperta articolata con richiesta di giustificazione.

L'item 23 risulta essere quello con un funzionamento maggiormente differenziato tra le due modalità di somministrazione, con uno scarto in termini di difficoltà pari a 2,38 *logit*. Nello stimolo di questo item è rappresentato un piano cartesiano e la domanda chiede di trovare le coordinate di uno dei punti caratteristici di una figura geometrica, di cui è fornita solo la descrizione verbale. In questo caso, molto probabilmente, la risoluzione su carta ha permesso agli studenti di disegnare la figura sul piano cartesiano per poi ricavare le coordinate del punto richiesto. L'impossibilità di interagire direttamente con la figura su schermo ha di fatto ristretto il ventaglio di strategie che gli studenti potevano mettere in atto per la risoluzione del problema in formato CBT.

Un problema analogo si è riproposto per quasi tutti gli item che presentavano grafici, istogrammi o figure geometriche di cui era richiesta più che la semplice lettura. Nella modalità CBT non risulta infatti possibile scrivere sulla figura, completarla con elementi utili alla comprensione o alla risoluzione, come punti, rette o

segmenti, prendere misure o evidenziare dati di rilievo. L'impossibilità di interagire con la figura inibisce dunque la messa in atto da parte dello studente di strategie di esplorazione e controllo, indispensabili nei quesiti relativi alla risoluzione di problemi (Schoenfeld, 1985; Zan, 2007). Poco praticabile è risultata altresì per gli studenti la strategia di riprodurre in tutto o in parte le figure richieste sui fogli messi a loro disposizione, per ragioni di tempo o di scarsa accuratezza nel disegno.

Gli item 36 e 47 richiedevano invece la scrittura di formule e simboli, che, nonostante le istruzioni fornite a tutti gli studenti per eseguire il compito, risulta usualmente più complessa in ambiente *computer based* che sulla carta.

In generale, osservando nel suo complesso la tabella 11, si può notare che gli item con un funzionamento notevolmente diverso nelle due modalità sono comunque di numero contenuto (12 su un totale di 54 item considerati).

Il grafico a dispersione (*scatter plot*) che segue riporta i valori stimati del parametro di difficoltà di ciascun item in entrambe le modalità di somministrazione: *computer based* lungo l'asse delle ascisse e cartacea lungo l'asse delle ordinate. In figura è riportato anche il grafico della bisettrice del primo e del terzo quadrante, sulla quale si trovano i punti corrispondenti ad item il cui parametro di difficoltà assume lo stesso valore in entrambe le modalità di somministrazione.

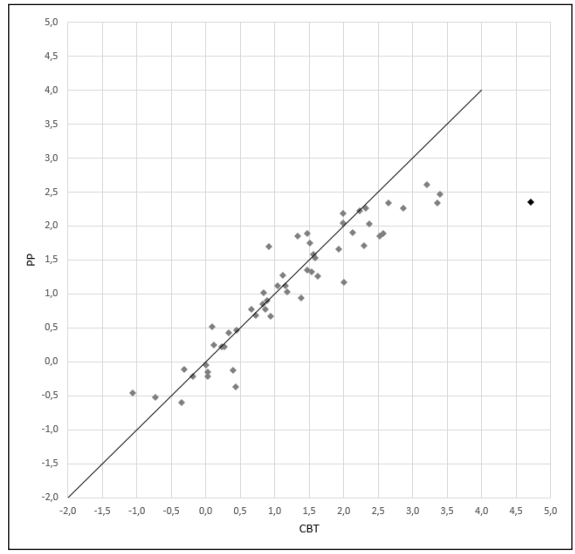

Grafico 1 – Item per valore del parametro di difficoltà *b*

Osservando il grafico 1 e la tabella 5 si nota che la maggior parte degli item mostra variazioni del parametro di difficoltà nel passaggio dalla modalità PPT alla modalità CBT, sia in positivo sia in negativo, sebbene sia evidente che vi è una prevalenza di item che risultano più difficili in modalità *computer based* (33 su 54). Un solo item si presenta come *outlier* (item 23).

Gli item analizzati costituiscono per contenuti, formato e difficoltà, una rappresentazione di tutti quelli realizzati fino ad oggi dall'INVALSI per le rilevazioni nazionali di matematica nella scuola secondaria di secondo grado. Poiché gli item sono gli stessi in ciascuna delle due modalità di somministrazione, è stato effettuato un *t* test per campioni accoppiati sulla differenza dei valori del parametro di difficoltà *b*.

Le medie e le deviazioni standard nei due gruppi di item, CBT e PPT, sono risultate molto simili e l'indice di correlazione fra le stime del parametro di difficoltà nelle due modalità è elevato: 0,92.

	Differenze accoppiate					t	gl	Sign. (a due code)
	Media	Deviazione std.	Errore standard della media	Intervallo di confidenza della differenza al 95%				
				Inferiore	Superiore			
CBT - PPT	0,167	0,486	0,066	0,034	0,299	2,519	53	0,015

Tabella 6 - Test per campioni accoppiati *(b)*

L'ipotesi che nella popolazione la differenza fra le medie sia nulla non può essere accettata (P-value < 0.05) e l'intervallo di confidenza al 95% contiene solo valori positivi. Gli item somministrati in versione CBT risultano quindi, in media, significativamente più difficili di quelli in formato cartaceo, sebbene tale differenza risulti comunque molto ridotta.

Se si limita l'analisi ai soli item a scelta multipla semplice o complessa (26) l'ipotesi che le medie siano uguali non può essere rifiutata (P-value = 0,711) e l'intervallo di confidenza al 95% è un intorno molto piccolo di zero (-0,09 ; 0,13), con media della differenza pari a 0,02. La correlazione dei parametri di difficoltà degli item è ancora molto elevata: 0,94.

Osservando i risultati per ciascun item, si vede che è presente un solo *outlier*, l'item 23, che dista 4,5 deviazioni standard dalla media. Cinque item distano più di una deviazione standard dalla media a sinistra (in CBT risultano più facili che in PPT), e sei risultano a una o più deviazioni standard dalla media a destra (in CBT risultano più difficili che in PPT).

La tabella che segue riassume la situazione.

Id	Item	Formato	CBT	PP	CBT - PP	Numero di Dev. Std.
7	A7	RG	0,922	1,695	-0,773	-1,9
20	B4	MCC	-1,052	-0,464	-0,588	-1,6
43	D2	RU	1,330	1,858	-0,528	-1,4
9	A9	RU	0,095	0,523	-0,428	-1,2
22	B6	MC	1,472	1,895	-0,423	-1,2
36	C6	RG	2,519	1,853	0,666	1,0
35	C5	RU	2,573	1,886	0,687	1,1
54	D13	MCC	0,439	-0,372	0,811	1,3
47	D6	RU	2,010	1,169	0,841	1,4
29	B13	RU	3,394	2,466	0,928	1,6
30	B14	RU	3,357	2,344	1,013	1,7
23	**B7**	**RU**	**4,724**	**2,349**	**2,375**	**4,5**

Tabella 7 - Item con variazioni significative del parametro di difficoltà *(b)*

7 Per ulteriori approfondimenti si veda il Rapporto tecnico delle Prove INVALSI 2016 (INVALSI, 2016), disponibile all'indirizzo: https://invalsi-areaprove.cineca.it/docs/file/002_Rapporto_tecnico_2016.pdf

Il test *t* per campioni accoppiati ripetuto eliminando l'item *outlier* (23), non fornisce risultati significativamente diversi da quello effettuato sull'intero insieme di item: P-value = 0,021, intervallo di confidenza al 95% non comprendente lo zero (0,02 ; 0,23), media delle differenze pari a 0,125 e correlazione fra i parametri di difficoltà uguale a circa 0,94.

2.3. *L'indice di discriminatività*

Il funzionamento delle domande è stato analizzato anche in termini di differenza dell'indice di discriminatività (*Id*), misurato mediante il calcolo del coefficiente di correlazione punto biseriale, coefficiente che può andare da 0 a 1 (Tab. 8). Per la valutazione delle differenze, è stato adottato il criterio utilizzato dall'INVALSI per la costruzione delle prove nazionali, che pone la soglia di accettabilità di un item a un livello di discriminatività di 0,20[7].

Id	Item	CBT	PPT	Formato		ID	Item	CBT	PPT	Formato
1	A1	0,44	0,39	MC		28	B12	0,39	0,52	CL
2	A2	0,51	0,45	RU		29	B13	0,41	0,34	RU
3	A3	0,01	0,01	MC		30	B14	0,38	0,32	RU
4	A4	0,06	0,03	MCC		**31**	**C1**	**0,20**	**0,38**	**MC**
5	A5	0,46	0,39	RU		32	C2	0,10	0,08	MC
6	A6	0,44	0,53	RU		33	C3	0,35	0,37	RU
7	A7	0,45	0,44	RG		34	C4	0,43	0,47	RU
8	**A8**	**0,09**	**0,42**	**RU**		35	C5	0,40	0,51	RU
9	A9	0,44	0,49	RU		36	C6	0,47	0,42	RG
10	A10	0,29	0,31	MCC		37	C7	0,11	0,17	MC
11	**A11**	**0,42**	**0,28**	**MC**		38	C8	0,52	0,50	RU
12	A12	0,28	0,21	MC		39	C9	0,27	0,31	MCC
13	A13	-0,10	0,06	MCC		40	C10	0,32	0,50	RU
14	A14	0,37	0,37	RG		41	C11	0,22	0,21	RU
15	**A15**	**0,06**	**0,23**	**MC**		42	D1	0,42	0,46	RU
16	A16	0,06	0,09	MC		43	D2	0,53	0,45	RU
17	B1	0,35	0,38	MC		44	D3	0,26	0,23	MC
18	B2	0,42	0,47	RG		45	D4	0,39	0,38	MCC
19	B3	0,49	0,48	RU		46	D5	0,40	0,45	RU
20	B4	0,22	0,29	MCC		47	D6	0,44	0,53	RU
21	B5	0,45	0,44	RU		48	D7	0,10	-0,01	MC
22	B6	0,41	0,48	MC		**49**	**D8**	**0,19**	**-0,01**	**MC**
23	**B7**	**0,00**	**0,46**	**RU**		50	D9	0,12	0,06	MC
24	B8	0,25	0,33	MCC		51	D10	0,14	0,04	MC
25	B9	0,33	0,51	RU		52	D11	0,19	0,18	MC
26	B10	0,54	0,53	RU		**53**	**D12**	**0,01**	**0,20**	**MC**
27	B11	0,31	0,50	RU		**54**	**D13**	**0,03**	**0,30**	**MCC**

Tabella 8 - Indice di discriminatività *(Id)*

Nel complesso gli item tendono a mantenere lo stesso potere discriminante nella maggior parte dei casi (38 item su 54). Quando vi sono variazioni di rilievo, sono prevalentemente a vantaggio della modalità cartacea (12 item). Solo 4 item migliorano in modo rilevante il loro indice di discriminatività nel passaggio alla modalità *computer based*.

Osservando la tabella 8, si notano tre item (8, 23 e 54) il cui potere discriminante è al di sotto della soglia di accettabilità nella versione CBT mentre risulta buono nella versione cartacea (item 8: *Id* = 0,09 in CBT e *Id* = 0,42 in PPT; item 23: *Id* = 0,00 in CBT e *Id* = 0,46 in PPT; item 54: *Id* = 0,03 in CBT e *Id* = 0,30 in PPT). Questi item contengono grafici o diagrammi non manipolabili in formato CBT oppure richiedono di utilizzare nelle risposte un linguaggio formale o simbolico.

Anche nel caso degli item 15 e 53 (entrambi a scelta multipla semplice) si osserva una insufficiente discriminatività nella versione CBT (item 15: *Id* = 0,06; item 53: *Id* = 0,01) rispetto a quella PPT (item 15: *Id* = 0,23; item 53: *Id* = 0,20). In questi casi, però, lo scarto in termini di capacità discriminante è molto più contenuto, passando dall'essere al di sotto della soglia in CBT all'essere appena al di sopra in PPT. Entrambi questi item presentavano nello stimolo figure geometriche da analizzare.

Nella versione CBT si rilevano differenze consistenti nei valori dell'indice di discriminatività rispetto alla versione PPT per alcuni item a scelta multipla semplice (11 e 49). L'item 49, seppur poco discriminante in entrambe le versioni, risulta esserlo maggiormente in quella CBT (*Id* = 0,19 in CBT e *Id* = - 0,01 in PPT). L'item 11, invece, abbastanza discriminante in PPT (*Id* = 0,28), mostra un funzionamento migliore in CBT (*Id* = 0,42).

2.4. *Le mancate risposte*

Nella tabella 9 sono riportate le percentuali di risposte omesse agli item nella versione PPT e nella versione CBT.

Nella fase preliminare di *data-cleaning* le mancate risposte sono state differenziate tra omesse (*missing*) e non raggiunte e, come si è detto, gli studenti che non avevano risposto a più di metà delle domande sono stati eliminati dal campione. In fase di analisi pertanto, essendo il numero di non raggiunte molto basso, è stato considerato solo il numero di risposte omesse.

In generale, si può osservare come nella modalità CBT il numero di risposte omesse diminuisca (per 34 item) o si mantenga costante (per 8 item), in particolare per gli item a risposta aperta univoca. Gli item che invece registrano un aumento rilevante nella percentuale di risposte omesse sono prevalentemente quelli a risposta aperta con richiesta di giustificazione.

In media la variazione della percentuale di *missing* nella versione CBT è pari a – 2,88, con una deviazione standard di 6,80.

Se si prendono in considerazione solo gli item per cui lo scarto in punti percentuali del numero di risposte omesse è in valore assoluto maggiore del 6%, si può notare come in questa categoria rientrino quasi esclusivamente gli item a risposta aperta, sia univoca sia articolata. Complessivamente 22 item su 54 presentano una variazione rilevante nel numero di risposte omesse e nella maggior parte dei casi (18) si osserva una diminuzione del loro numero nella modalità *computer based*. Gli item 48 e 49 sono gli unici a scelta multipla semplice per i quali si osserva una differenza nel numero di risposte omesse superiore al 6% in valore assoluto.

Id	CBT	PPT	CBT - PPT	Formato		Id	CBT	PPT	CBT - PPT	Formato
47	**41,30**	**56,59**	**-15,29**	**RU**		23	43,04	45,86	-2,82	RU
34	**13,81**	**28,74**	**-14,93**	**RU**		15	10,28	12,70	-2,42	MC
9	**17,67**	**32,55**	**-14,88**	**RU**		13	0,00	2,00	-2,00	MCC
27	**45,00**	**59,05**	**-14,05**	**RU**		52	7,83	9,44	-1,61	MC
38	**21,76**	**34,01**	**-12,25**	**RU**		8	30,60	32,16	-1,56	RU
2	**0,00**	**12,16**	**-12,16**	**RU**		31	2,93	4,45	-1,52	MC
35	**14,23**	**25,51**	**-11,28**	**RU**		22	14,29	15,36	-1,07	MC
28	**21,05**	**30,97**	**-9,92**	**CL**		40	78,99	79,27	-0,28	RU
33	**14,64**	**23,89**	**-9,25**	**RU**		4	0,00	0,00	0,00	MCC
21	**18,97**	**28,09**	**-9,12**	**RU**		10	0,00	0,00	0,00	MCC
42	**19,57**	**26,92**	**-7,35**	**RU**		20	0,00	0,00	0,00	MCC
48	**4,80**	**12,15**	**-7,35**	**MC**		24	0,00	0,00	0,00	MCC
49	**10,92**	**18,23**	**-7,31**	**MC**		45	0,00	0,00	0,00	MCC
26	**35,92**	**43,10**	**-7,18**	**RU**		54	0,00	0,00	0,00	MCC
25	**30,37**	**37,36**	**-6,99**	**RU**		17	3,88	3,75	0,13	MC
43	**56,09**	**62,64**	**-6,55**	**RU**		44	4,35	3,85	0,50	MC
5	**6,47**	**12,94**	**-6,47**	**RU**		11	1,33	0,78	0,55	MC
41	**30,34**	**36,48**	**-6,14**	**RU**		51	5,41	4,44	0,97	MC
46	25,65	30,22	-4,57	RU		30	8,67	7,11	1,56	RU
3	2,59	7,06	-4,47	MC		1	4,74	2,75	1,99	MC
19	27,59	31,84	-4,25	RU		16	3,77	1,68	2,09	MC
32	7,53	11,74	-4,21	MC		7	28,88	25,10	3,78	RG
53	7,44	11,11	-3,67	MC		6	44,83	39,61	5,22	RU
39	15,90	19,51	-3,61	MCC		**29**	**46,74**	**38,53**	**8,21**	**RU**
37	8,37	11,74	-3,37	MC		**36**	**41,42**	**27,53**	**13,89**	**RG**
12	6,33	9,49	-3,16	MC		**18**	**42,24**	**27,72**	**14,52**	**RG**
50	3,57	6,67	-3,10	MC		**14**	**55,76**	**38,31**	**17,45**	**RG**

Tabella 9 - Numero di risposte omesse in %

2.5. *Le differenze nella stima delle abilità*

Per rispondere alla seconda domanda della ricerca e verificare se gli studenti ottengano risultati diversi in relazione alla modalità con cui hanno sostenuto la prova, ne è stata stimata l'abilità secondo il modello di Rasch a 1 parametro a partire da ciascuna coppia di prove costituite dagli stessi item, e quindi di fatto equivalenti dal punto di vista del contenuto matematico.

Preliminarmente, si è proceduto ad ancorare internamente tra loro le prove T1. Esse contengono infatti 5 item comuni di ancoraggio, provenienti da prove utilizzate nelle rilevazioni dell'INVALSI degli anni scorsi e dunque somministrati a tutti gli studenti italiani di seconda superiore. Tali item, scelti sulla base delle loro caratteristiche psicometriche, sono stati vincolati ad assumere i valori dei parametri misurati sul campione nazionale INVALSI. Gli item non di ancoraggio delle prove T1 CBT sono stati invece vincolati ai parametri degli item delle corri-

spondenti prove T1 PPT. Considerato che i due gruppi che hanno sostenuto le prove nelle due modalità di somministrazione sono equivalenti in termini di abilità (vedi paragrafo 1.2), vincolando gli item delle prove *computer based* ai valori delle corrispondenti prove cartacee, eventuali differenze nella stima delle abilità sono da attribuire interamente alla modalità di somministrazione.

Sui due gruppi di studenti CBT e PPT è stato effettuato un test *t* per campioni indipendenti sull'uguaglianza delle medie, a varianze uguali presunte, per verificare se le stime delle abilità risultassero uguali.

Le statistiche generali evidenziano che le medie e le deviazioni standard nei due gruppi sono sostanzialmente identiche e i risultati del test *t* mostrano, come atteso, che non è possibile rifiutare l'ipotesi che le medie delle abilità siano uguali nella popolazione, con una probabilità del 95%:

Gruppo	N	Media	Deviazione std.	Errore standard della media
CBT	933	-0,127	0,976	0,032
PPT	951	-0,104	1,112	0,036

Tabella 10 - Statistiche di gruppo – Abilità T1

Test t per l'eguaglianza delle medie						
t	gl	Sign. (a due code)	Differenza della media	Errore standard della differenza	Intervallo di confidenza della differenza al 95%	
					Inferiore	Superiore
-0,469	1882	0,639	-0,023	0,048	-0,117	0,072

Tabella 11 - Test per campioni indipendenti – Abilità T1

3. Conclusioni

Le analisi condotte nel presente contributo hanno cercato di indagare se esistano, e in tal caso quale entità abbiano, differenze nel funzionamento degli item nel passaggio dalla somministrazione in formato cartaceo alla somministrazione *computer based*.

Nonostante l'INVALSI per l'implementazione delle prove in CBT abbia proceduto secondo un modello di migrazione delle domande volto a minimizzare le differenze fra le due modalità di somministrazione, l'analisi ha messo in luce la presenza di variazioni nel parametro di difficoltà degli item, nell'indice di discriminatività e nel numero di risposte omesse.

Prendendo in considerazione il parametro *b* di difficoltà, si osserva che la prova è risultata mediamente più difficile in formato *computer based.* Le differenze tra i singoli item risultano, nella maggior parte dei casi, di ridotta entità, mentre sono rilevanti per 12 item su 54.

Nei casi in cui si sono riscontrate variazioni significative nel parametro *b* di difficoltà, queste riguardavano principalmente item che sono risultati più difficili in formato CBT. Tali item hanno in comune la caratteristica di presupporre un certo grado di interazione del soggetto con il quesito, poiché richiedevano di scri-

vere formule o simboli o la manipolazione delle immagini. Per questo tipo di quesiti si può ipotizzare che la somministrazione *computer based* abbia ristretto il campo delle strategie che lo studente poteva mettere in atto per rispondere. In questo senso, quindi, sembrano venire rafforzate le osservazioni di Bennett (Bennett et al. 2008) sul fatto che la somministrazione informatizzata possa inibire la possibilità di utilizzare strategie di soluzione alternative dei quesiti.

Gli item risultati invece significativamente più facili in CBT appartenevano a quesiti a grappolo, costituiti da uno stimolo iniziale a cui erano collegati più item, ad ognuno dei quali era attribuito un distinto punteggio. Si può quindi ipotizzare che il fatto che lo stimolo, nella modalità CBT, fosse ripetuto per ciascun item e che gli item fossero visualizzati separatamente, abbia rafforzato l'indipendenza fra l'uno e l'altro e facilitato la riflessione su ognuno di essi singolarmente considerato.

Si osserva inoltre che, nel complesso, gli item tendono a mantenere lo stesso potere discriminante nella maggior parte dei casi (38 item su 54) e che le variazioni significative sono prevalentemente a vantaggio della modalità cartacea (12 item su 16).

Infine, un risultato che di per sé sembra essere incoraggiante è la consistente diminuzione delle risposte omesse nella modalità CBT, soprattutto nel caso delle domande aperte. Tale fenomeno può essere interpretato anche alla luce di quanto sostenuto in letteratura (Richardson et al., 2002) secondo cui la risoluzione di problemi in CBT può essere percepita come più stimolante e motivante, a prescindere dall'effettiva difficoltà del quesito posto.

Le analisi fin qui condotte forniscono spunti di riflessione meritevoli di essere ulteriormente approfonditi.

In primo luogo sarà interessante verificare se le differenze in termini di difficoltà degli item riguardino allo stesso modo tutti gli studenti o se esse pesino in modo diverso su particolari gruppi (studenti con abilità alte/basse, maschi/femmine, etc.).

Inoltre, risulterà utile analizzare i risultati del questionario sulla familiarità con l'uso del computer, sia per indagare la correlazione tra il punteggio nella prova cognitiva e le competenze informatiche, sia per valutare se la familiarità con l'uso del computer sia, e in che misura, un predittore dei risultati nel test *computer based*.

Riferimenti bibliografici

Bennett R. E., Braswell J., Oranje A., Sandene B., Kaplan B., Yan F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. U.S.A. *The Journal of Technology, Learning, and Assessment*, *6*(9), 4-38.

Bennett R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5-25.

Bridgeman B., Lennon M. L., Jackenthal A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16*, 191-205.

INVALSI (2016). *Rilevazioni nazionali degli apprendimenti 2015 – 2016. Rapporto tecnico.* Estratto da http://www.invalsi.it/invalsi/doc_evidenza/2016/002_Rapporto_tecnico_2016.pdf

JISC (2006). *e-Assessment Glossary*. Estratto da https://www.webarchive.org.uk/wayback/archive/20140615085353/http://www.jisc.ac.uk/media/documents/themes/elearning/eassess_glossary_extendedv101.pdf

Johnson M., Green S. (2004). *On-line assessment: the impact of mode on student performance.,* U.K. Cambridge: Paper presented at the British Educational Research Association Annual Conference, University of Manchester, 16-18 September 2004.

Johnson M., Green S. (2006). On-Line Mathematics Assessment: The Impact of Mode on Performance and Question Answering Strategies. *Journal of Technology, Learning, and Assessment, 4*(5).
Karkee T., Kim D., Fatica K. (2010). *Comparability Study of Online and Paper and Pencil Tests Using Modified Internally and Externally Matched Criteria.* Paper presented at the annual meeting of the American Educational Research Association (AERA), Denver, CO, April 29 – May 4, 2010
Thompson N. A., Weiss D. J. (2009). Computerized and Adaptive Testing in Educational Assessment. In F. Scheuermann, J. Björnsson (Eds.), *The Transition to Computer Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing* (pp. 127-133). European Commission: IPSC.
Pepper D. (2012). *KeyCoNet 2012 Literature Review: Assessment for Key competences*. Estratto da http://www.obrazovanje.org/rs/uploaded/dokumenta/keyconet-2013-literature-review_-assessment-for-key-competences.pdf
OECD (2012). *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, OECD Publishing.
Pisa (2015a). *Draft Mathematics Framework, 2013.* Estratto da https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Mathematics%20Framework%20.pdf
Pisa (2015b). *Field Trial Goals, Assessment Design and Analysis Plan For Cognitive Assessment, 2014*. Estratto da http://www.oecd.org/callsfortenders/Annex%20F1_FTAnalysisPlan-Cognitive_1.pdf
Poggio J., Glasnapp D. R., Yang X., Poggio A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*(6).
Pommerich M. (2004). Developing computerized versions of paperand-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment,* 2(6).
Richardson M., Baird J. A., Ridgway J., Ripley M., Shorrocks-Taylor D., Swan M. (2002). Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving. *Computer in Human Behavior*, *18*(6), 633-649
Ripley M. (2009). Transformational Computer-based Testing. In F. Scheuermann, J. Björnsson (Eds.), *The Transition to Computer Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing* (pp. 92-98). European Commission: IPSC.
Schoenfeld A. (1985). *Mathematical Problem Solving*. New York: Academic Press.
Wang S., Jiao H., Young M.J., Brooks T.E., Olson J. (2007). A meta-analysis of testing mode effects in Grade K—12 mathematics tests. *Educational and Psychological Measurement*, *67*, 219-238.
Zan R. (2007). *Difficoltà in matematica. Osservare, interpretare, intervenire.* Milano: Springer-Verlag Italia.