Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione

WORKING PAPER N. 49/2020

La procedura automatizzata per la codifica delle domande a risposta aperta delle prove INVALSI in modalità CBT

Michele Marsili - INVALSI

Emiliano Campodifiori - INVALSI

Cecilia Bagnarol - INVALSI

Silvia Donno - INVALSI

Collana: Working Papers INVALSI

ISSN: 2611 - 5719

Abstract

Il presente lavoro di ricerca illustra la nuova procedura di correzione automatizzata delle domande a risposta aperta, introdotte per l'a.s. 2018-19, per le prove INVALSI di Italiano e Matematica somministrate in modalità CBT (Computer based Test) agli studenti dei gradi 8 (terza secondaria di primo grado), 10 (seconda secondaria di secondo grado) e 13 (quinta secondaria di secondo grado). Il team INVALSI, costituito da statistici e informatici e impegnato nella codifica delle risposte aperte, ha implementato un algoritmo per il trattamento di stringhe di testo più o meno articolate.

L'approccio metodologico utilizzato, da annoverarsi tra i metodi di correzione automatizzata supervisionata, rappresenta un valido compromesso tra una codifica manuale e una totalmente automatizzata tipica degli algoritmi di Machine Learning. L'utilizzo di questa metodologia, infatti, ha il vantaggio di ridurre sensibilmente le ore/uomo necessarie allo svolgimento della correzione degli item a risposta aperta, se paragonata a una codifica manuale, e di acquisire una maggiore precisione riducendo le occorrenze di codifiche errate, se paragonata alla codifica automatizzata non supervisionata.

Parole chiave: correzione automatica supervisionata; risposte aperte; codifica manuale; algoritmo; CBT.

Keywords: automatic supervised correction; open-ended short answers; manual correction; algorithm; CBT.

Introduzione

La valutazione rappresenta una fase cruciale del processo educativo (Berry, 2003; Cucchiarelli et al., 2000), dal momento che consente di verificare e di accertare le conoscenze acquisite dagli studenti.

Nel contesto tipico di un esame o di una verifica, la valutazione implica che il docente o esaminatore fornisca allo studente un feedback alle risposte da lui fornite ad un set di domande relative all'argomento del test. Vi sono tuttavia alcune circostanze in cui la presenza dell'esaminatore non è disponibile ma lo studente ha comunque necessità di ricevere una valutazione delle proprie conoscenze (Mohler e Mihalcea, 2009), in queste circostanze si fa spesso uso di prove computerizzate.

Il D.Lgs. 62/2017 ha introdotto nuove norme in materia di valutazione e certificazione delle competenze nel sistema scolastico italiano. Una delle più importanti novità riguarda l'introduzione di prove standardizzate Computer Based (CBT) per le rilevazioni nazionali degli apprendimenti per gli studenti della terza secondaria di primo grado, della seconda e quinta secondaria di secondo grado. La modalità CBT consiste nella somministrazione di una prova valutativa in forma computerizzata, quindi non più svolta con l'utilizzo di carta e penna ma con monitor e tastiera (INVALSI, 2019). Negli ultimi anni le grandi innovazioni tecnologiche hanno permesso una sempre maggiore implementazione di questa modalità nel sistema educativo: in ambito internazionale OECD (Organisation for Economic Co-operation and Development) ha ufficialmente introdotto la modalità CBT per le prove di apprendimento a partire dall'indagine PISA 2015, allo stesso modo IEA (International Association for the Evaluation of Educational Achievement) con le indagini ePIRLS del 2016 e TIMMS del 2019.

A partire dall'a.s. 2017-2018, anche INVALSI ha effettuato, per le classi menzionate dal D.Lgs. 62/2017, il passaggio da una prova cartacea ad una prova CBT, svolta cioè con l'ausilio del computer.

Il passaggio dai test somministrati in modalità cartacea a quelli in modalità CBT ha permesso di utilizzare un range sempre più ampio di tipologie di domande per indagare più a fondo le competenze degli studenti (Scheuermann & Björnsson, 2009; Valenti et al., 2000; Parshall et al., 2000). Tuttavia, l'introduzione della nuova modalità di somministrazione delle prove "computer based" e l'implementazione di item sempre più sofisticati e complessi pone inevitabilmente l'attenzione sulla valutazione automatica delle prove (*Automatic Assessment*), e di conseguenza sulla correzione, o codifica, centralizzata e automatizzata delle risposte degli studenti, in special modo per gli item che prevedono una risposta aperta.

Comprensibilmente, se le domande che richiedono risposte a scelta multipla sono più facilmente valutabili con una procedura automatizzata e metodi computazionali, le domande a risposta aperta (*open-ended short answers*) necessitano di una tecnologia in grado di valutare il linguaggio naturale (*natural language*) oppure

il testo strutturato della notazione matematica (Burrows et al., 2014). Per questo tipo di item, infatti, non sono sempre posti a priori vincoli sul numero o sulla tipologia di caratteri - o di parole - digitabili dallo studente (ad esempio numeri e/o caratteri speciali), allo scopo di lasciare allo studente la libertà di esprimersi simulando il più possibile l'approccio della prova cartacea.

Si rende così necessario lo sviluppo di una procedura di correzione ad hoc dotata di un certo grado di automatizzazione che, mantenendo un accettabile livello di accuratezza e precisione nella codifica, permetta di superare i limiti di ore-lavoro a disposizione. Il grado di automatizzazione ottenuto mediante l'implementazione di una procedura di correzione può intuitivamente ritenersi inversamente correlato al fabbisogno di ore-lavoro e alla precisione attesa per la procedura stessa. Si consideri da un lato la correzione classica o "*manuale*", che massimizza la precisione attesa ma anche il fabbisogno di ore-lavoro e dall'altro lato la correzione totalmente automatica svolta da uno specifico algoritmo, che consente di minimizzare le ore-lavoro ma determina prevedibilmente una maggiore frequenza di errori di correzione: tra questi due estremi si inserisce la procedura implementata nel 2018 dal team INVALSI che potremmo definire "automatica supervisionata" e che realizza un buon compromesso nel trade-off fra precisione e fabbisogno ore-lavoro.

La descrizione della procedura è riportata di seguito applicata alla codifica delle prove CBT di Italiano e Matematica dell'a.s. 2018-19 delle classi terze della scuola secondaria di primo grado (di seguito Grado 8) e delle classi quinte della scuola secondaria di secondo grado (Grado 13).

Dati e Metodi

Le prove INVALSI sono state somministrate a tutti gli studenti (*prove censuarie*) delle classi seconde della scuola secondaria di primo e delle classi quinte della scuola secondaria di secondo grado, in modalità CBT (Computer Based Test), per Italiano, Matematica e Inglese (Listening e Reading). Tale modalità prevede che la prova di ciascuno studente si componga di domande estratte da un ampio repertorio di quesiti (*banca di item*) e vari dunque da studente a studente, mantenendo tuttavia per ciascuna prova la medesima difficoltà e struttura.

Nelle prove somministrate con fascicolo cartaceo il compito di codificare le risposte agli item a risposta aperta era affidato ai docenti degli alunni delle classi interessate e/o agli osservatori INVALSI che di fatto correggevano manualmente le risposte aperte. Con l'avvento della prova "computer based" questa modalità di codifica diviene difficilmente percorribile se non inapplicabile, si è resa quindi necessaria la

centralizzazione della codifica delle domande a risposta aperta e lo sviluppo di una procedura di correzione automatizzata.

Durante la somministrazione delle prove INVALSI, ogni computer utilizzato dallo studente è connesso ad un server principale adibito alla raccolta delle prove degli studenti, per tutte le materie, in un arco temporale predefinito: l'insieme delle prove di Italiano, Matematica e Inglese costituisce la base dati oggetto di correzione e codifica.

Il team INVALSI ha implementato una procedura di codifica delle risposte fornite dagli studenti ottenendo un algoritmo per il trattamento di stringhe di testo più o meno articolate. Tale procedura prevede lo svolgimento di alcune importanti operazioni propedeutiche che potremmo definire "fase di collaudo" per il team di correzione:

- il team si confronta con i gruppi degli autori degli item di ciascuna materia (Italiano, Matematica, Inglese) per definire in modo chiaro e univoco i criteri di correzione, ovvero un insieme di regole in base alle quali è possibile discriminare le risposte corrette da quelle errate in riferimento a ciascun item oggetto di somministrazione;
- il team traduce i criteri di correzione definiti con gli autori in pattern logico-informatici (definiti "*espressioni regolari*" o "*Regex*", di cui parleremo in seguito) interpretabili dal computer e utili a classificare in modo del tutto automatico le risposte processate;
- il team procede al collaudo vero e proprio dell'algoritmo verificando, per le risposte fornite dagli studenti durante la fase di pre-test della prova cognitiva, il grado di concordanza tra la codifica prodotta mediante la correzione manuale e quella elaborata automaticamente dall'algoritmo implementato. L'algoritmo e i pattern logico-informatici che traducono le condizioni di correzione sono ritenuti sufficientemente accurati e allineati alle indicazioni di codifica degli autori quando vi è perfetta convergenza tra le due codifiche;
- il team pianifica insieme agli autori una serie di attività di controllo, produzione di reportistica e ricezione di feedback da parte degli autori da svolgersi durante il periodo di somministrazione delle prove. Tali attività hanno lo scopo di verificare l'accuratezza della codifica in produzione e, qualora gli autori lo ritengano necessario, modificare in corsa le condizioni di correzione, sia per semplici affinamenti che per modifiche significative.

Ultimata la fase di collaudo della procedura, prende avvio l'insieme di operazioni che realizzano di fatto la codifica automatica delle risposte aperte in quattro passaggi distinti:

1. *acquisizione del database* contenente le risposte oggetto di codifica e riferito ad un preciso arco temporale di somministrazione;

2. *operazioni di "data-cleaning"*: le risposte sono sottoposte alla correzione automatica degli errori di battitura e/o ortografici e sono private di tutti gli elementi considerati non utili alla correzione (segni di interpunzione, parentesi ecc.).
3. *operazioni di codifica*: l'algoritmo confronta ciascuna risposta con la versione logica delle condizioni di correzione (le "espressioni regolari" o RegEx) classificandola in corretta o errata;
4. *produzione di un report* relativo alla codifica in corso avente funzione di fornire un quadro chiaro sulla classificazione delle risposte da parte dell'algoritmo.

Nella prima fase si acquisisce il database da processare e si effettuano controlli mirati per evidenziare eventuali anomalie nei dati e nella loro struttura.

Durante la fase di *data cleaning* prende avvio la manipolazione effettiva delle risposte. Tale fase consiste in una serie di operazioni sulle stringhe di testo (le risposte) allo scopo di eliminare gli elementi invarianti ai fini della classificazione, riducendo, di fatto, la variabilità delle modalità di risposta possibili e dunque la complessità del set da classificare. Più in dettaglio, le operazioni a cui generalmente sono sottoposte le stringhe di testo prevedono:

- identificazione e rimozione di punteggiatura, caratteri speciali articoli preposizioni e congiunzioni;
- lemmatizzazione delle parole, ossia la riduzione di una forma flessa di una parola alla sua forma canonica;
- correzione automatica degli errori di battitura e ortografia, ossia individuazione di parole "fuori vocabolario" e sostituzione delle stesse con la corrispondente versione corretta.

Nella fase successiva il database, "normalizzato" dopo il data-cleaning, viene processato da un algoritmo che esegue la classificazione di ciascuna stringa in corretta o errata mediante l'utilizzo di un motore "*RegEX*", un software capace di confrontare stringhe di testo con la versione logica delle condizioni di correzione chiamata "*espressione regolare*" o "*RegEx*", predisposta in fase di collaudo. In sostanza, un'*espressione regolare* è una sequenza di caratteri che identifica univocamente un insieme anche non finito di stringhe. Tradurre in espressione regolare una condizione di correzione significa definire l'insieme possibile di risposte corrette ad uno specifico item, pertanto il motore RegEx processa le stringhe di risposta e classifica come corrette solo quelle che appartengono all'insieme precedentemente definito dalla RegEx.

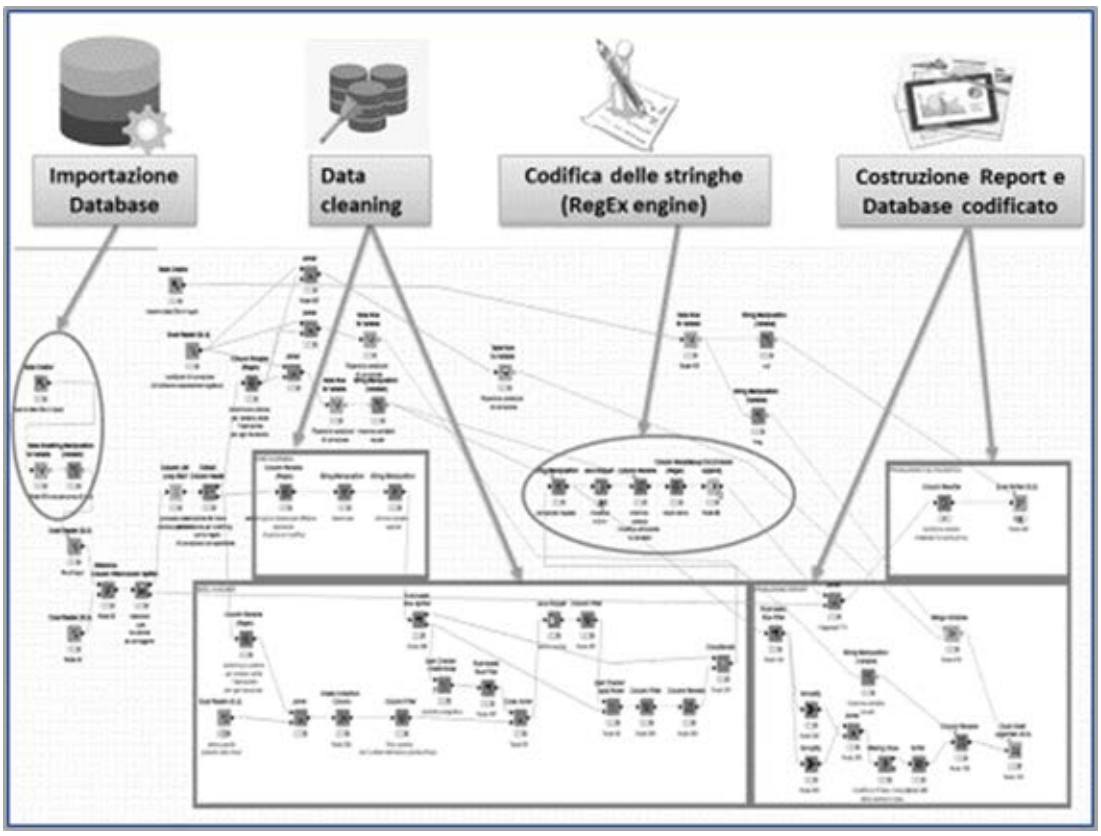
La fase conclusiva prevede la produzione, ad intervalli concordati con gli autori durante tutta la finestra di somministrazione, di report relativi alla correzione. Lo scopo dei report è fornire uno strumento per valutare la qualità di codifica dell'algoritmo e per evidenziare eventuali errori di classificazione; ciascun report fornisce la distribuzione di frequenza delle risposte fornite dagli studenti e la relativa classificazione assegnata, per ogni intervallo di tempo considerato. I principali destinatari dei report sono i gruppi di lavoro degli autori a

cui è consentito di verificare l'esigenza di introdurre variazioni alle condizioni di correzione al fine di migliorare l'accuratezza e la precisione della classificazione prodotta.

L'insieme di tutte le operazioni che costituiscono la procedura di codifica sinora illustrata sono state implementate mediante la piattaforma analitica open source KNIME Analytics Platform (Berthold et al., 2008).

Tale software è dotato di una interfaccia grafica mediante la quale si possono organizzare e collegare fra di loro le unità di computazione base chiamate "nodi" al fine di crearne un insieme strutturato chiamato "workflow", un flusso analitico che, nel caso specifico, mette in pratica le operazioni costituenti la procedura di codifica desiderata. Ogni nodo svolge un compito ben preciso (avviare un ciclo, caricare dati, trasformare dati, eseguire funzioni ecc.) ed equivale quindi ad un passo del processo analitico che si vuole realizzare. Il workflow è un insieme di nodi opportunamente ordinati e collegati fra di loro "in cascata". Il posizionamento reciproco dei nodi definisce il loro ordine di esecuzione: il workflow, quindi, può essere avviato mandando in esecuzione il primo nodo (ossia la prima operazione del "flusso di lavoro") che, terminato il suo compito, determina in automatico l'avvio del nodo successivo (la seconda operazione del "flusso di lavoro") e così via fino al completamento dell'intero flusso analitico (De Mauro, 2019) (Figura 1).

Figura 1 - Workflow della codifica delle risposte aperte



Risultati

Il core della procedura di codifica descritta in precedenza è chiaramente l'insieme delle espressioni regolari associate al gruppo di item che costituiscono una prova. Una espressione regolare deve essere costruita in modo da poter intercettare, per ciascun item, sia la risposta corretta sia tutte le più probabili varianti della stessa, considerate altresì corrette dagli autori: a titolo di esempio, in Figura 2 è riportato un esempio di Regex (nel riquadro superiore) che per costruzione riesce ad individuare (nel riquadro inferiore) non solo la risposta corretta "ufficiale" (in questo caso 17/2) ma anche tutta una serie di probabili varianti corrette della stessa (8,5; 8.5; 17 fratto 2). L'aspetto più complesso nella costruzione delle espressioni regolari è proprio quello di dotarle di una accettabile sensibilità discriminatoria.

Figura 2 – Esempio di traduzione di una condizione di correzione in espressione regolare

La qualità e flessibilità di ciascuna espressione regolare determina la qualità generale della codifica prodotta: un insieme di espressioni regolari poco flessibili aumenterebbe significativamente il numero di falsi negativi, ossia di tutte le varianti alla chiave di risposta corretta non riconosciute dalla Regex. Considerando il numero elevato di prove oggetto di codifica durante una rilevazione nazionale, l'eventualità che nelle risposte degli studenti si presentino un numero importante di varianti delle chiavi di correzione è uno scenario piuttosto verosimile; a conferma di ciò, sono indicative le tabelle a seguire (Tabelle 1, 2, 3, 4) da cui si può evincere

l'elevato numero di varianti della chiave di correzione che le Regex costruite sono riuscite a codificare correttamente.

Tabella 1 - Prova di Italiano, Main Study 2019, Grado 8.

Item	Tipo di interazione	N. di varianti della chiave di
Item 1	Extended text	3942
Item 2	Extended text	3289
Item 3	Extended text	1810
Item 4	Text entry	1619
Item 5	Block + Text entry	1298
Item 6	Extended text	1175
Item 7	Block + Text entry	927
Item 8	Text entry	714
Item 9	Extended text	673
Item 10	Extended text	541
Item 11	Block + Text entry	492
Item 12	Block + Text entry	342
Item 13	Text entry	291
Item 14	Extended text	286
Item 15	Extended text	256
Item 16	Block + Text entry	232

Tabella 2 - Prova di Matematica, Main Study 2019, Grado 8.

Item	Tipo di interazione	N. di varianti della chiave di
Item 1	Text entry	650
Item 2	Text entry	381
Item 3	Block + Text entry	375
Item 4	Text entry	350
Item 5	Block + Text entry	322
Item 6	Text entry	317
Item 7	Text entry	306
Item 8	Block + Text entry	184
Item 9	Block + Text entry	155
Item 10	Text entry	119
Item 11	Block + Text entry	112
Item 12	Text entry	96
Item 13	Text entry	94
Item 14	Block + Text entry	93
Item 15	Text entry	87
Item 16	Text entry	76
Item 17	Block + Text entry	75
Item 18	Block + Text entry	73

Tabella 3 - Prova di Italiano, Main Study 2019, Grado 13.

Item	Tipo di interazione	N. di varianti della chiave
Item 1	Extended text	968
Item 2	Extended text	591
Item 3	Extended text	587
Item 4	Extended text	420
Item 5	Extended text	294
Item 6	Extended text	216
Item 7	Extended text	211
Item 8	Extended text	209
Item 9	Text entry	154
Item 10	Text entry	111
Item 11	Text entry	86
Item 12	Extended text	84
Item 13	Text entry	83
Item 14	Text entry	83
Item 15	Text entry	80
Item 16	Extended text	71
Item 17	Text entry	65
Item 18	Text entry	54

Tabella 4 - Prova di Matematica, Main Study 2019, Grado 13.

Item	Tipo di interazione	N. di varianti della chiave
Item 1	Text entry	1416
Item 2	Text entry	761
Item 3	Text entry	631
Item 4	Text entry	320
Item 5	Text entry	159
Item 6	Text entry	157
Item 7	Text entry	148
Item 8	Text entry	141
Item 9	Text entry	94
Item 10	Text entry	86
Item 11	Text entry	79
Item 12	Text entry	76
Item 13	Text entry	74
Item 14	Text entry	70

Conclusioni

L'introduzione nel sistema scolastico di prove standardizzate Computer based per la rilevazione delle competenze degli studenti ha reso indispensabile e imprescindibile l'adattamento della correzione delle prove ad un approccio automatizzato. Il metodo di correzione automatizzata supervisionata degli item a risposta aperta adottato da INVALSI ha richiesto, nelle diverse fasi di implementazione della procedura, un assiduo confronto tra team di correzione, costituito da statistici e informatici, e i gruppi di autori di ciascuna materia (Italiano, Matematica, Inglese). Questo lungo e complesso lavoro, tuttavia, ha concesso di ottenere un algoritmo che garantisce, con una precisione prossima al 100%, una corretta classificazione delle risposte degli studenti. Il fine ultimo della codifica automatica, infatti, è quello di rilasciare a scuole e studenti una corretta certificazione delle competenze, scevra da errori di classificazione.

Risulta chiaro, tuttavia, che se comparato a una procedura di correzione completamente automatizzata, come quelle ottenibili con gli algoritmi ML (machine learning), la procedura automatizzata supervisionata richiede un maggior numero di ore di lavoro e coinvolge un maggior numero di risorse.

L'adozione di buone pratiche, il miglioramento delle procedure di pre-processing e la creazione di una base dati che contenga una raccolta di migliaia di tipologie di risposte degli studenti, potrebbero determinare una notevole riduzione dell'ammontare delle ore di lavoro.

Il metodo adottato dal team di correzione INVALSI può quindi rappresentare un buon compromesso tra la codifica manuale e l'adozione degli algoritmi ML per il raggiungimento di performance che massimizzino la precisione della codifica.

D'altro canto, sarebbe opportuno non tralasciare e portare avanti in futuro una riflessione tra team di correzione e gruppi di esperti delle diverse discipline per valutare l'adozione di modelli completamente automatizzati che, se ben ideati e predisposti, potrebbero riservare in futuro buone garanzie per performance puntuali, efficaci e accurate.

Bibliografia

Berthold M.R. et al. (2008), KNIME: The Konstanz Information Miner. In: Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R. (eds) *Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Berlin, Heidelberg

Berry R. (2003), *Alternative assessment and assessment for learning", in Proceedings of the 29th IAEA Conference, theme*: Societies Goals and Assessment, USA.

Cucchiarelli A., Panti M., Valenti S. (2000), "Web-based assessment of Student Learning. In A.K. Aaggarwal (ed) *Web-Based Learning and Teaching Technologies: Opportunities and Challenges*, 175-197, Idea Group Publishing.

De Mauro A. (2019), *Big data analytics. Analizzare e interpretare dati con il machine learning*, Apogeo.

Gazzetta Ufficiale della Repubblica Italiana. *DECRETO LEGISLATIVO 13 aprile 2017, n. 62*. Text available: https://www.gazzettaufficiale.it/eli/id/2017/05/16/17G00070/sg, retrieved: 6 marzo 2019.

INVALSI (2019), "*Le prove Computer Based (CBT). Terzo anno scuola secondaria di I grado (Grado 8) a.s. 2019-2020. Organizzazione delle prove CBT*". Text available on: https://invalsi-areaprove.cineca.it/docs/2020/2019-2020_Organizzazione%20delle%20prove%20CBT_Grado_08.pdf, retrieved: 8 marzo 2020.

Magliano J.P., Graesser A.C. (2012), Computer-based assessment of student-constructed responses, *Behavior Research Methods* 44(3):608-21.

Molher M., Mihalcea R., (2009), Text-to-Text Semantic Similarity for Automatic Short Answer Grading, Conference: EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, March 30 - April3, 2009, Athens, Greece.

Scheuermann F., Björnsson J. (2009), *The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, Office for Official Publications of the European Communitites, Luxembourg.

Parshall C.G., Davey T., Pashley P.J. (2000), Innovative Item Types for Computerized Testing, in: van der Linden W.J., Glas G.A. (eds) *Computerized Adaptive Testing: Theory and Practice*. Springer, Dordrecht.