RATE

Reasoning through Adaptive Test Evaluation

Abstract Numerical Verbal

Guida all'uso

GIUNTI PSYCHOMETRICS

INDICE

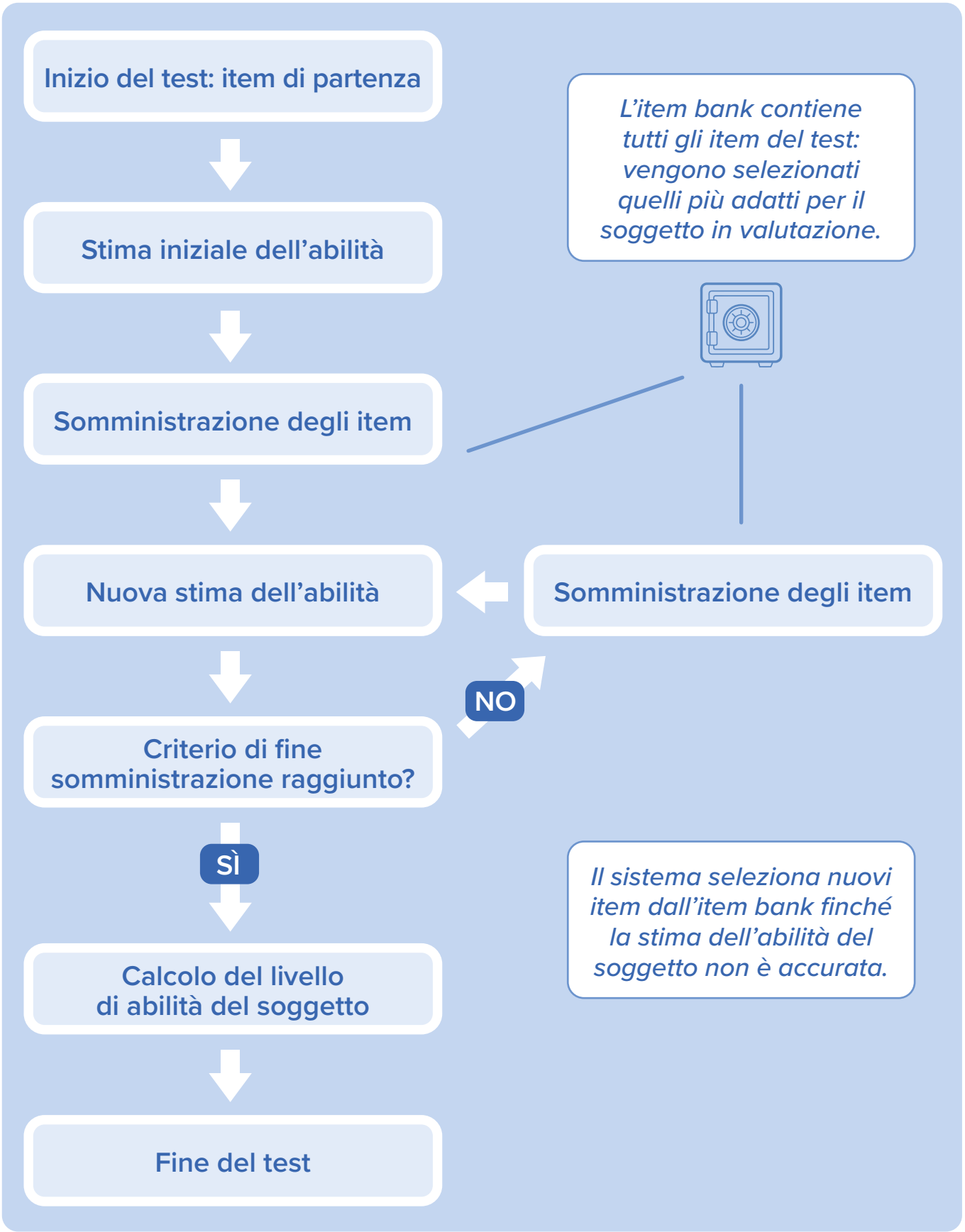
INTRODUZIONE AI TEST ADATTIVI

I test adattivi computerizzati (Computerized Adaptive Testing; CAT), in particolare il *Reasoning through Adaptive Test Evaluation* (RATE), rappresentano una delle innovazioni più significative nella valutazione psicometrica moderna. A differenza dei test tradizionali a somministrazione fissa, che propongono lo stesso insieme di item a tutti i partecipanti, i test adattivi selezionano dinamicamente gli item da somministrare in base alle risposte date in precedenza dal soggetto. Questo meccanismo consente al test di adattarsi in tempo reale al livello di abilità stimato della persona, presentando item più facili in caso di risposte errate e più difficili in caso di risposte corrette.

Tale peculiarità garantisce numerosi vantaggi. In primo luogo, permette di ottenere stime più rapide e precise delle abilità individuali, riducendo il numero complessivo di item necessari e, di conseguenza, diminuendo la durata del test. Ciò rende l'esperienza della valutazione meno faticosa e più efficiente, con elevati standard di accuratezza. Inoltre, i test adattivi ottimizzano l'uso del pool di item e accrescono la sicurezza del materiale: poiché i percorsi seguiti dai partecipanti sono unici, si riduce significativamente la possibilità di esposizione sistematica degli item e, quindi, il rischio di compromissione del test.

Dal punto di vista psicometrico, i test adattivi si basano su modelli di teoria della risposta all'item (Item Response Theory, IRT), che consentono di stimare con precisione sia la difficoltà degli item sia l'abilità del soggetto. L'integrazione tra algoritmi di selezione, criteri di arresto e stime aggiornate in tempo reale rende i CAT strumenti potenti, flessibili e altamente personalizzabili per diversi contesti di valutazione, dalla psicologia clinica alla selezione del personale, dall'educazione alla ricerca scientifica.

Fasi principali del testing adattivo computerizzato (CAT)*



* Vedi anche il paragrafo "Implementazione del Computerized Adaptive Testing (CAT)".

CREAZIONE DEL REASONING THROUGH ADAPTIVE TEST EVALUATION (RATE)

La serie di nuovi test adattivi è composta da tre differenti strumenti: il *Reasoning through Adaptive Test Evaluation – Abstract* (RATE-A), il *Reasoning through Adaptive Test Evaluation – Numerical* (RATE-N) e il *Reasoning through Adaptive Test Evaluation – Verbal* (RATE-V). Questi test sono finalizzati a misurare l'abilità nelle aree di ragionamento astratto, ragionamento numerico e ragionamento verbale in persone con diploma di scuola secondaria superiore o laurea.

La rilevanza delle abilità di ragionamento nella previsione della performance lavorativa è stata ampiamente dimostrata da numerosi studi presenti nella letteratura scientifica, sia a livello nazionale sia internazionale (Hartigan e Wigdor, 1989; Schmidt e Hunter, 2004). Le abilità di ragionamento sono in tal senso risultate tra le variabili che esercitano la maggiore influenza sulla performance lavorativa (Schmidt, Hunter e Outerbridge, 1986).

I test sviluppati sono destinati alla selezione e alla valutazione di candidati per posizioni lavorative di diverso livello, con l'obiettivo di identificare le capacità di problem solving avanzate e di gestione complessa delle informazioni. Questi test sono stati progettati in modo da non misurare le competenze o le capacità acquisite durante il percorso di studi. Infatti, le conoscenze scolastiche hanno solo un ruolo marginale nella risoluzione delle diverse prove.

RATE-A

Si colloca all'interno del vasto campo degli strumenti sviluppati per misurare ciò che è chiamato, con diverse etichette, "abilità di ragionamento astratto", "intelligenza fluida" o "pensiero divergente". Indipendentemente dal termine utilizzato, questo test è composto da stimoli figurativi che valutano la capacità del soggetto di riconoscere nuovi schemi, ideare nuovi

metodi e operare a vari livelli di analisi mentale. Il test fornisce una misura dell'abilità di identificare rapidamente regole logiche in un insieme di elementi nuovi e non familiari, integrando queste informazioni per risolvere i problemi proposti.

RATE-N

Valuta la capacità di pensare in modo efficace con i numeri. È importante notare che le prove di ragionamento numerico non devono essere confuse con i test di rendimento, che misurano le conoscenze matematiche del soggetto in valutazione piuttosto che la sua abilità di ragionamento. Gli stimoli sono stati progettati per esaminare la capacità di pensare metodicamente, formulare giudizi pertinenti, selezionare informazioni rilevanti e ragionare in modo efficiente.

RATE-V

È progettato per valutare la capacità di ragionamento verbale dei candidati, ovvero la capacità di dedurre e valutare il significato e la logica di contenuti verbali, ragionare per concetti ed esplicitarne le relazioni.

Nella **Figura 1** sono riportati degli esempi di item, uno per tipologia di test: RATE-A, RATE-N e RATE-V.

RACCOLTA DATI

I test RATE-A, RATE-N e RATE-V sono stati implementati su una piattaforma online finalizzata a ottenere una raccolta dati standardizzata. L'item bank iniziale era composto da 466 item per il test di ragionamento verbale, 404 per il test di ragionamento astratto e 450 per il test di ragionamento numerico. La somministrazione degli item è avvenuta in modalità randomizzata, con l'inclusione di 20 item selezionati sulla base delle loro caratteristiche psicometriche e utilizzati come ancoraggi.

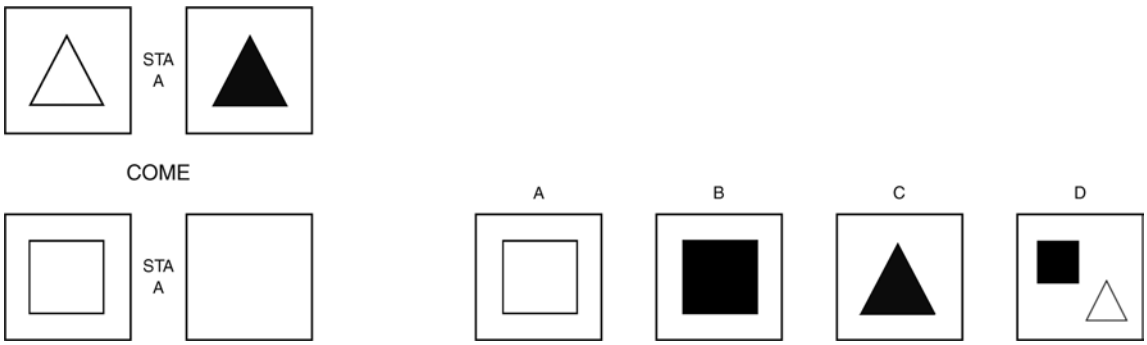
Ogni sessione prevedeva la somministrazione di 35 item, con l'obiettivo di ridurre il carico cognitivo sul soggetto e mantenere la comparabilità con i dati di somministrazione standard. Sono stati inoltre inseriti 3 item attentivi per monitorare eventuali cali di attenzione ed escludere coloro che non rispettavano i criteri minimi di qualità del dato.

La raccolta dati è stata condotta su un campione di oltre 1100 soggetti distribuiti su tutto il territorio nazionale. Tutti

Figura 1. Esempi di item tratti dai test RATE-A, RATE-N e RATE-V

RATE-A

Qual è la figura mancante che completa in modo corretto la proporzione?



RATE-N

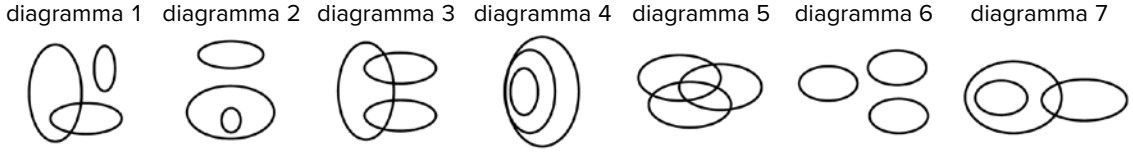
Per ciascun problema, ci sono quattro serie di numeri, due seguite da “sì” e due seguite da “no”. Nelle due serie seguite da “sì” la relazione che lega i numeri è la stessa, mentre nelle due serie seguite da “no” tale relazione non è presente.
Osserva la serie di numeri sulla sinistra. Identifica la relazione tra i numeri nelle due serie contrassegnate da “sì”. Quindi, scegli tra le serie sulla destra quella che segue la stessa relazione.

1	3	5	sì
3	4	5	no
2	5	8	no
2	4	6	sì

1	**2**	**3**
2	**4**	**9**
7	**9**	**11**

RATE-V

Scegli tra questi diagrammi quale, secondo te, rappresenta correttamente la relazione insiemistica esistente tra i termini presentati.



Tavolo, automobile, canoa.

gli altri item, al di fuori di quelli di ancoraggio e controllo, sono stati presentati in modalità casuale, garantendo una distribuzione omogenea delle occorrenze tra gli item, con lo scopo di raccogliere almeno 100 osservazioni per ciascun item.

ANALISI DESCRITTIVE

L'analisi è iniziata con la codifica delle variabili demografiche e la preparazione dei dati. Ciò ha incluso la gestione accurata di queste variabili e l'applicazione di criteri di esclusione basati sui tempi di risposta e sull'accuratezza delle risposte agli item di controllo dell'attenzione. In particolare, sono stati esclusi dal dataset il 5% dei soggetti con tempi di risposta più rapidi, oltre a coloro che non hanno superato uno o più dei 3 item attentivi somministrati. Successivamente, sono state calcolate le statistiche descrittive per i dataset ottenuti. Le analisi hanno riguardato tutte le variabili, inclusi i singoli item del test, considerando sia l'intero campione sia i sottogruppi definiti dai diversi livelli di istruzione (**Tabella 1**).

Tabella 1. Statistiche descrittive dei campioni

	RATE-A	RATE-N	RATE-V
N	1376	1174	1636
Età (DS)	25.41 (6.78)	25.24 (3.18)	25.15 (3.83)
Genere (M-F)			
M	53%	53%	50%
F	47%	47%	50%
Titolo di studio			
Diploma di scuola superiore o inferiore	42%	41%	42%
Laurea o superiore	58%	59%	58%
Area di residenza			
Nord-est	24%	23%	23%
Nord-ovest	27%	27%	27%
Centro	22%	23%	22%
Sud e Isole	27%	27%	28%

ANALISI IRT

Per ciascun dataset – ovvero RATE-A, RATE-N e RATE-V – è stato applicato un framework di Item Response Theory (IRT). L'analisi è iniziata con la valutazione della dimensionalità dei dati delle risposte agli item, che rappresenta una fase essenziale per verificare l'assunzione di unidimensionalità su cui si fonda la modellazione IRT. A tal fine è stato utilizzato il metodo NOHARM (Nonlinear Harmonic Analysis Robust Method), specificamente progettato per l'Analisi Fattoriale Esplorativa (AFE) di dati categorici o ordinali. Il metodo consente di stimare la struttura latente sottostante alle risposte attraverso un modello armonico non lineare e di valutarne l'adeguatezza tramite l'analisi dei residui.

Tabella 2. NOHARM: residui minimi e massimi dei 3 test

Test	Min	Max
RATE-A	–.09	.08
RATE-N	–.07	.07
RATE-V	–.06	.09

Per ciascun test è stata quindi esaminata la matrice dei residui standardizzati, utilizzata come indicatore della bontà dell'adattamento di una soluzione unidimensionale. I risultati confermano che per ciascun dataset la struttura dei dati è compatibile con un modello unidimensionale, giustificando l'applicazione di modelli IRT a un fattore latente (**Tabella 2**).

Dopo la valutazione della dimensionalità, sono state calcolate le statistiche descrittive a livello di item per fornire approfondimenti dettagliati sulle loro proprietà. I modelli IRT iniziali sono stati calibrati utilizzando sia le specifiche Rasch sia 2PL. Quando entrambi i modelli sono convergenti, sono state condotte analisi comparative di adattamento del modello. Alla fine, il modello Rasch è stato selezionato per tutti i dataset in ragione del miglior adattamento di tale modello, dell'interpretabilità e della parsimonia.

Poiché non tutti i soggetti hanno completato la stessa versione del test, sono stati incorporati item di ancoraggio con parametri predefiniti, al fine di migliorare l'accuratezza del mo-

dello e garantire il collegamento tra i dataset. Questi item di ancoraggio hanno stabilizzato la stima dei parametri durante le calibrazioni Rasch, garantendo la comparabilità tra i dataset.

La qualità dei modelli è stata valutata attraverso l'analisi delle statistiche di adattamento degli item, come gli indici di infit, e mediante la generazione di grafici caratteristici del test, tra cui i grafici di traccia, di informazione e di attendibilità (**Figure 2 e 3**). Sono stati inoltre calcolati i punteggi fattoriali ed è stata esaminata la precisione delle stime del tratto latente, l'analisi dell'attendibilità marginale ed empirica. Dopo la valutazione del modello, l'analisi è proseguita con la selezione finale degli item e il test del Differential Item Functioning (DIF).

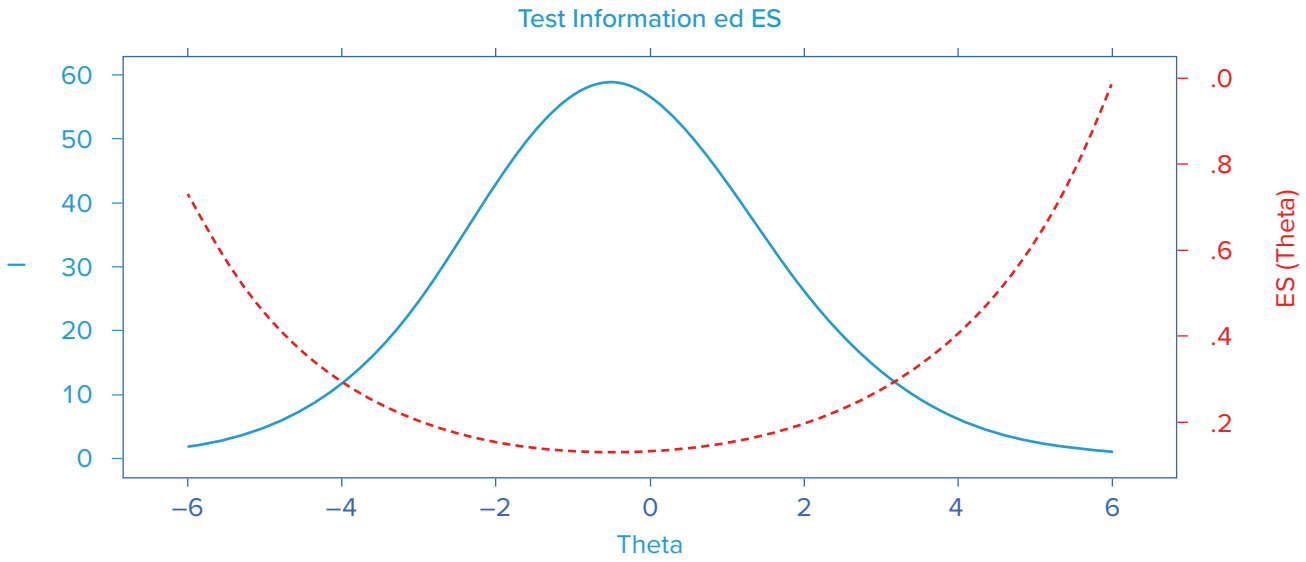
Le valutazioni di adattamento sono state eseguite seguendo le raccomandazioni di Linacre (2002), ciò significa che gli item con statistiche di adattamento *mean-square* tra .5 e 1.5 sono stati considerati idonei per la misurazione. Inoltre, gli item con valori theta (θ) inferiori a –3 o superiori a +3 *DS* sono stati considerati troppo facili o troppo difficili per lo scopo previsto della valutazione. Per questo motivo, sono stati classificati come fuori dai limiti ed esclusi dal pool.

I modelli Rasch finali sono stati quindi calibrati con gli item di ancoraggio fissati, consentendo un ulteriore affinamento dei parametri degli item. Le analisi del Differential Item Functioning (DIF) sono state successivamente condotte utilizzando sia i test del Rapporto di Verosimiglianza che il metodo Mantel–Haenszel, con correzione di continuità e senza purificazione degli item. Le dimensioni dell'effetto sono state stimate utilizzando la scala ETS Delta. Queste analisi hanno confrontato le prestazioni tra i gruppi con diverso titolo di studio per identificare eventuali item che mostravano un funzionamento differenziale.

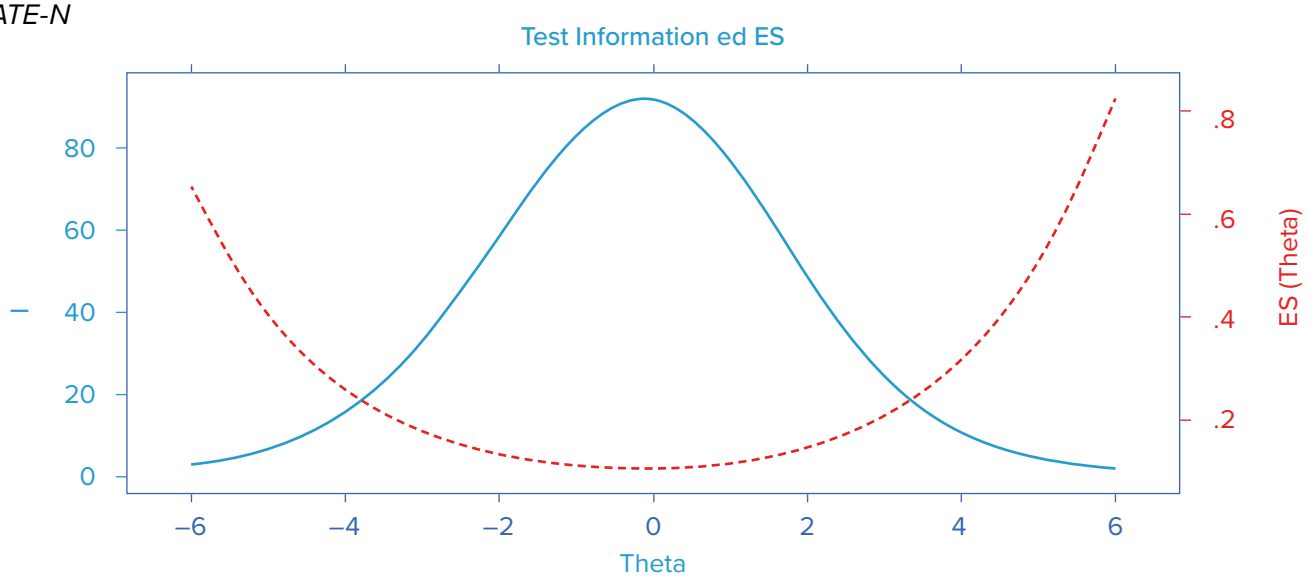
Il nuovo test adattivo è stato progettato per essere utilizzabile sia da diplomati che da laureati, senza una divisione per livello di educazione. Questo è stato possibile grazie all'analisi del Differential Item Functioning (DIF), che ha dimostrato che non ci sono differenze significative nel funzionamento degli item tra i due gruppi.

Figura 2. Test Information Function (I) ed errore standard (ES) di Theta

RATE-A



RATE-N



RATE-V

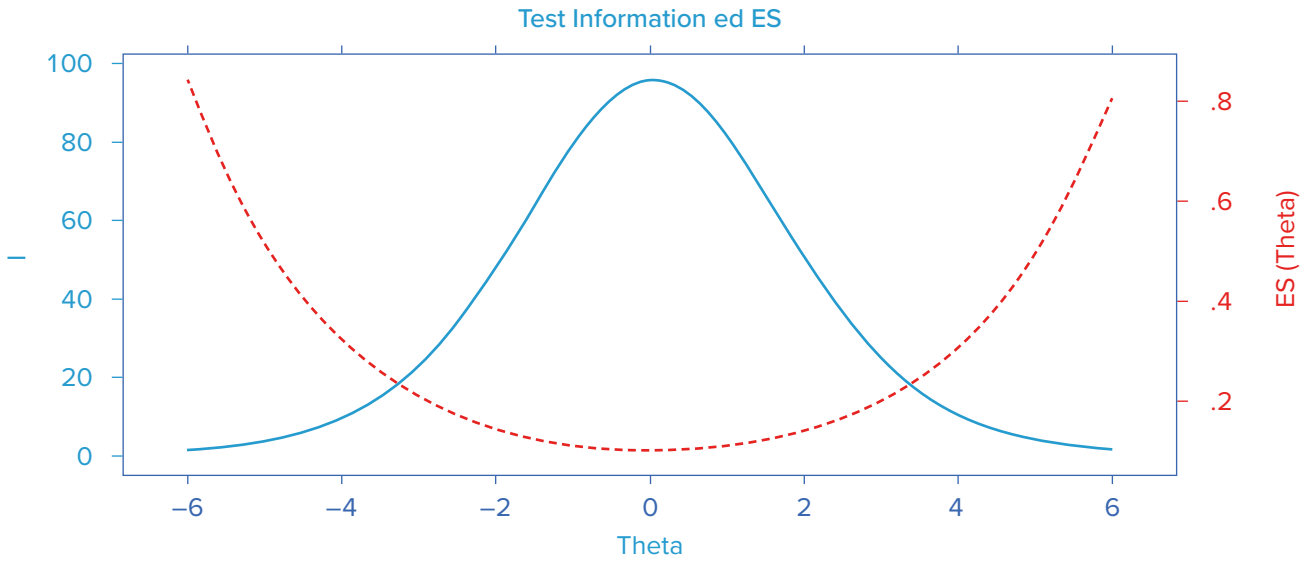
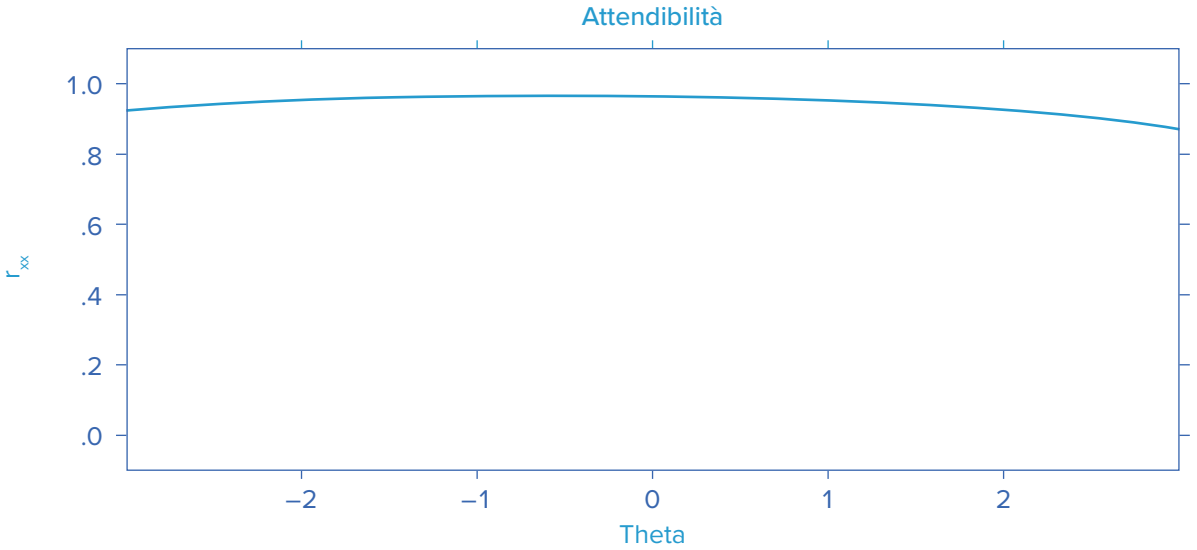
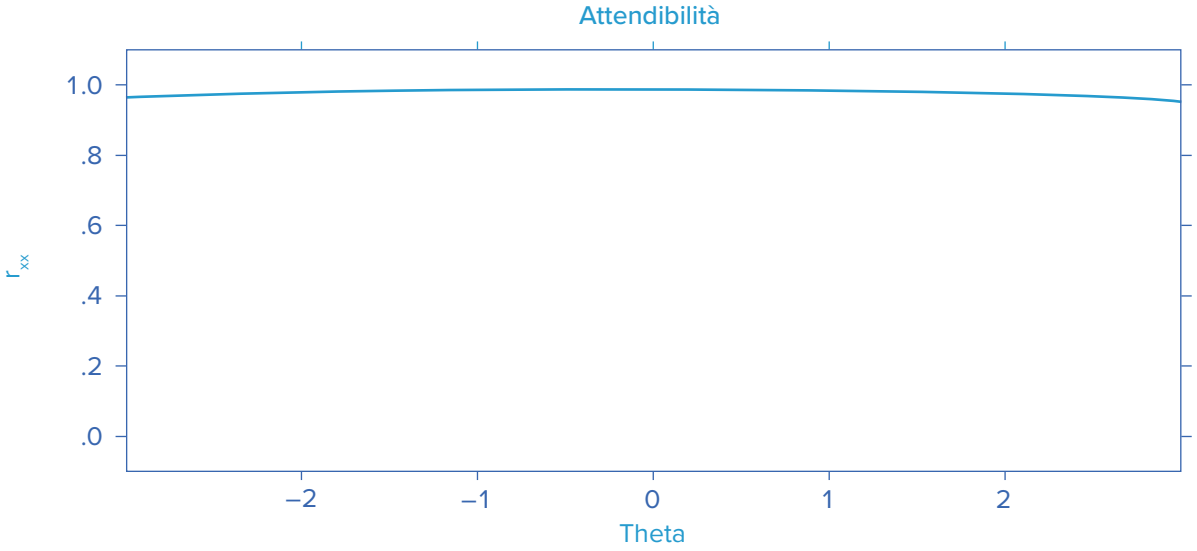


Figura 3. Grafici di attendibilità dei 3 test

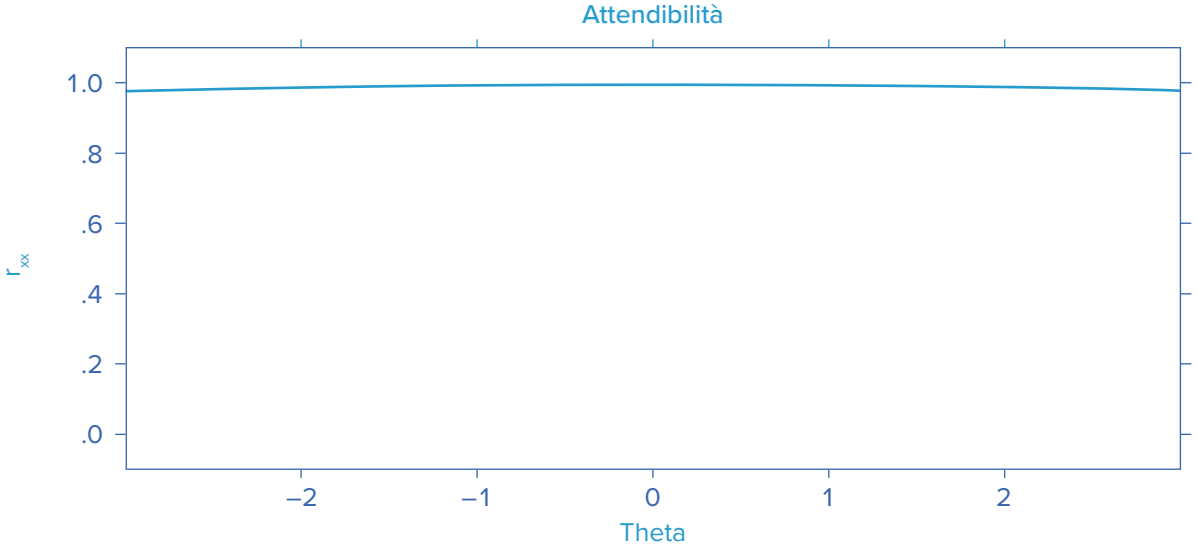
Attendibilità RATE-A



Attendibilità RATE-N



Attendibilità RATE-V



IMPLEMENTAZIONE DEL COMPUTERIZED ADAPTIVE TESTING (CAT)

A partire dai parametri degli item stimati durante la fase di modellazione, è stato sviluppato e implementato un sistema di Computerized Adaptive Testing (CAT) specifico per ciascuna delle tre versioni previste: RATE-A, RATE-N e RATE-V. Ciascun sistema CAT è stato costruito per massimizzare l'efficienza della misurazione, riducendo il numero di item somministrati e migliorando la precisione della stima dell'abilità individuale.

Dopo ogni risposta fornita dal soggetto, la stima dell'abilità è aggiornata utilizzando l'approccio bayesiano Expected A Posteriori (EAP), ovvero tenendo conto dell'informazione fornita dall'item a cui la persona ha appena risposto. Questo approccio consente di ottenere stime più stabili e robuste, soprattutto nelle fasi iniziali del test o in presenza di pattern di risposta "estremi" (per esempio, tutte le risposte sono corrette o tutte errate), che rappresentano situazioni in cui l'approccio tradizionale basato sulla massima verosimiglianza (Maximum Likelihood Estimation, MLE) tende a produrre stime meno affidabili o non definite. L'utilizzo della distribuzione a priori permette infatti di "ancorare" la stima, evitando derive non realistiche dovute alla scarsità delle informazioni durante la somministrazione dei primi item.

All'avvio del test, al soggetto in valutazione è attribuito un livello iniziale di abilità pari a zero (Theta = 0), in linea con l'assunzione di una distribuzione a priori normale standard ($\mu = 0$, $\sigma = 1$). Sono quindi somministrati 2 item iniziali selezionati in modo casuale tra quelli disponibili nella item bank, appartenenti a un sottoinsieme predefinito di item "di partenza". Questa fase ha lo scopo di raccogliere dati preliminari utili a una prima stima del livello di abilità del soggetto e dell'errore standard associato (ES), evitando nel contempo una sovraesposizione degli item iniziali.

Gli item di partenza per la somministrazione del test adattivo sono stati selezionati in base ai loro valori theta intermedi e al numero totale di item. Per quanto riguarda quest'ultimo criterio, è stato deciso di utilizzare circa un terzo del pool totale di item per evitare bias dovuti al tasso di esposizione e ad altri problemi che potrebbero sorgere da un set limitato di item

iniziali. Pertanto, per il RATE-A sono stati scelti 102 item con valori theta compresi tra –.5 e +.5 *DS*. Per il RATE-N sono stati selezionati 155 item all'interno dello stesso intervallo theta (–.5 e .5). Per il RATE-V sono stati selezionati 140 item con valori theta compresi tra –.3 e +.3 *DS*.

Successivamente, la somministrazione procede in modo adattivo: a ogni nuova risposta, la stima di abilità (theta) viene aggiornata utilizzando la regola di Bayes, tenendo conto dell'informazione fornita dall'item a cui la persona ha appena risposto. Sulla base della stima aggiornata, viene selezionato l'item successivo che massimizza l'informazione in corrispondenza del livello attuale di abilità del soggetto, secondo la funzione di informazione dell'item derivata dal modello IRT.

Questo meccanismo garantisce che ogni item somministrato sia ottimamente calibrato rispetto al livello stimato del soggetto, migliorando l'efficienza del test (minor numero di item per una stima altrettanto precisa) e riducendo l'onere cognitivo.

La procedura continua fino al soddisfacimento di almeno uno dei criteri di interruzione predefiniti: il raggiungimento di un errore standard inferiore a .40, oppure la somministrazione di un massimo di 40 item. Per garantire un livello minimo di informazione e robustezza della stima, è stato inoltre fissato un criterio di completamento minimo, che richiede la somministrazione di almeno 10 item prima di valutare eventuali condizioni di arresto. Indipendentemente da ciò, in nessun caso la durata delle somministrazioni può superare i 30 minuti.

Al termine della procedura di somministrazione adattiva, il livello di abilità finale stimato viene trasformato in un punteggio standard, utilizzando la distribuzione dei punteggi del campione normativo di riferimento. Questa trasformazione in punti T consente di facilitare l'interpretazione dei risultati da parte dei clinici e degli operatori, offrendo una metrica stabile e comparabile con altri strumenti psicometrici standardizzati.

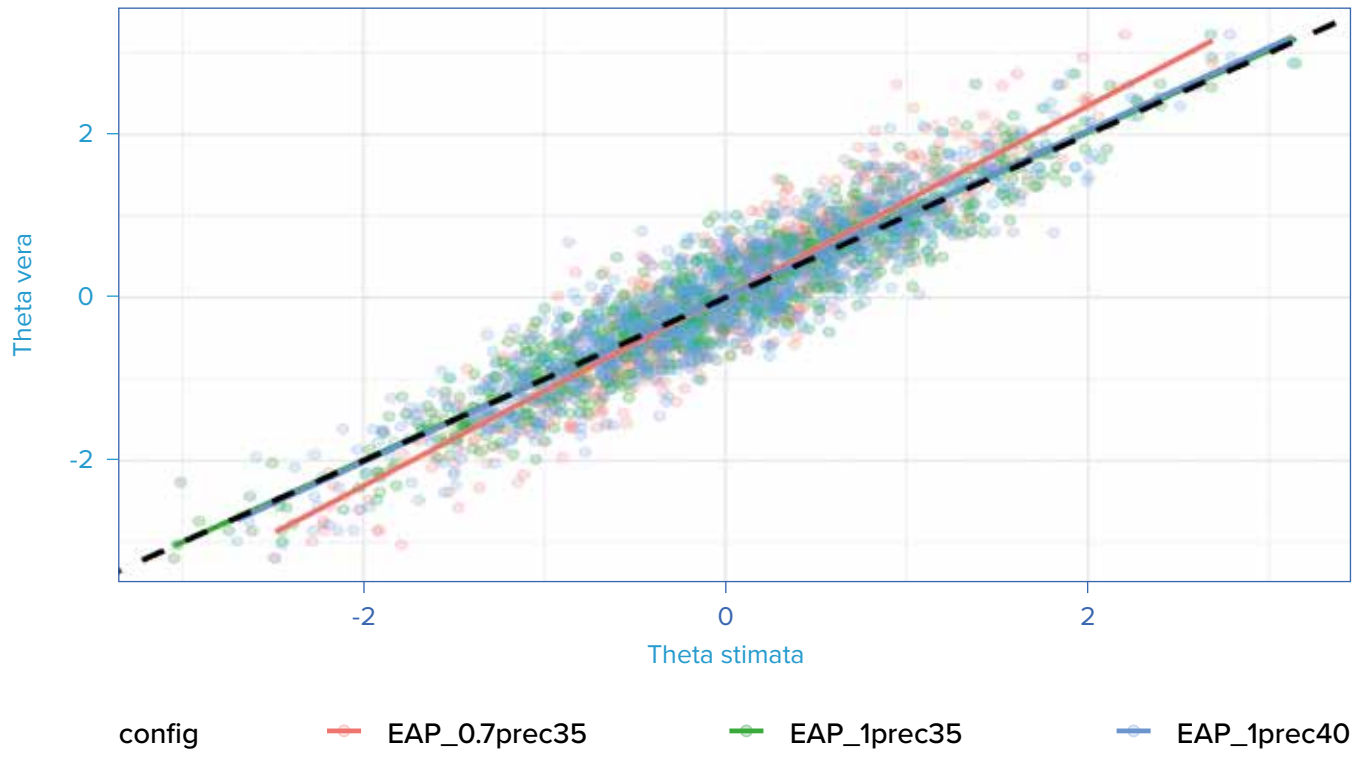
Sono anche state condotte una serie di simulazioni sui test adattivi per testarne la funzionalità, comparando tre differenti settaggi di parametri con lo scopo di trovare la combinazione migliore (**Figure 4, 5 e 6**). Il primo considera per la stima

dell'abilità dei *prior* con $M = 0$ e $DS = 1$ e un criterio di stop a 40 item o ES inferiore a .35 (EAP_1prec35), il secondo *prior* con $M = 0$ e $DS = .7$ e un criterio di stop a 40 item o ES inferiore a .35 (EAP_0.7prec35), mentre il terzo *prior* con $M = 0$ e $DS = 1$ e un criterio di stop a 40 item o ES inferiore a .40 (EAP_1prec40).

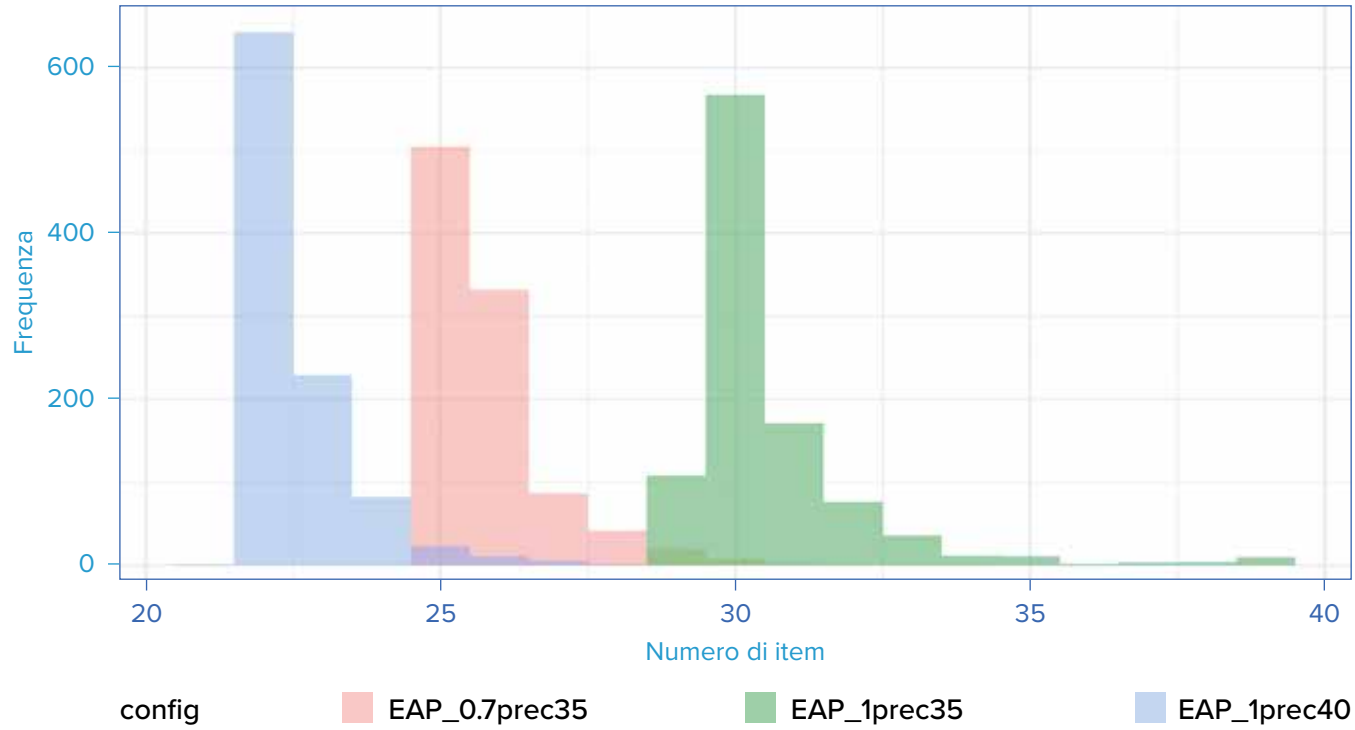
Come è possibile osservare dai risultati che simulano le risposte al CAT di 1000 soggetti aventi un'abilità (theta) distribuita normalmente, il settaggio con *prior* con $M = 0$ e $DS = 1$ e un criterio di stop a 40 item o ES inferiore a .40 fornisce il miglior bilanciamento tra affidabilità delle stime del livello di abilità (correlazioni .91-.93; ES medio .40-.41) e numero di item somministrati (n. item medio 22.5-29.4) nei tre test. Infatti, la configurazione EAP_1prec35, per quanto affidabile tende a produrre test eccessivamente lunghi, mentre la configurazione EAP_0.7prec35 consente la riduzione della lunghezza del test a discapito di una peggiore stima dei livelli di abilità più estremi.

Figura 4. RATE-A: simulazione delle risposte con diversi parametri di settaggio

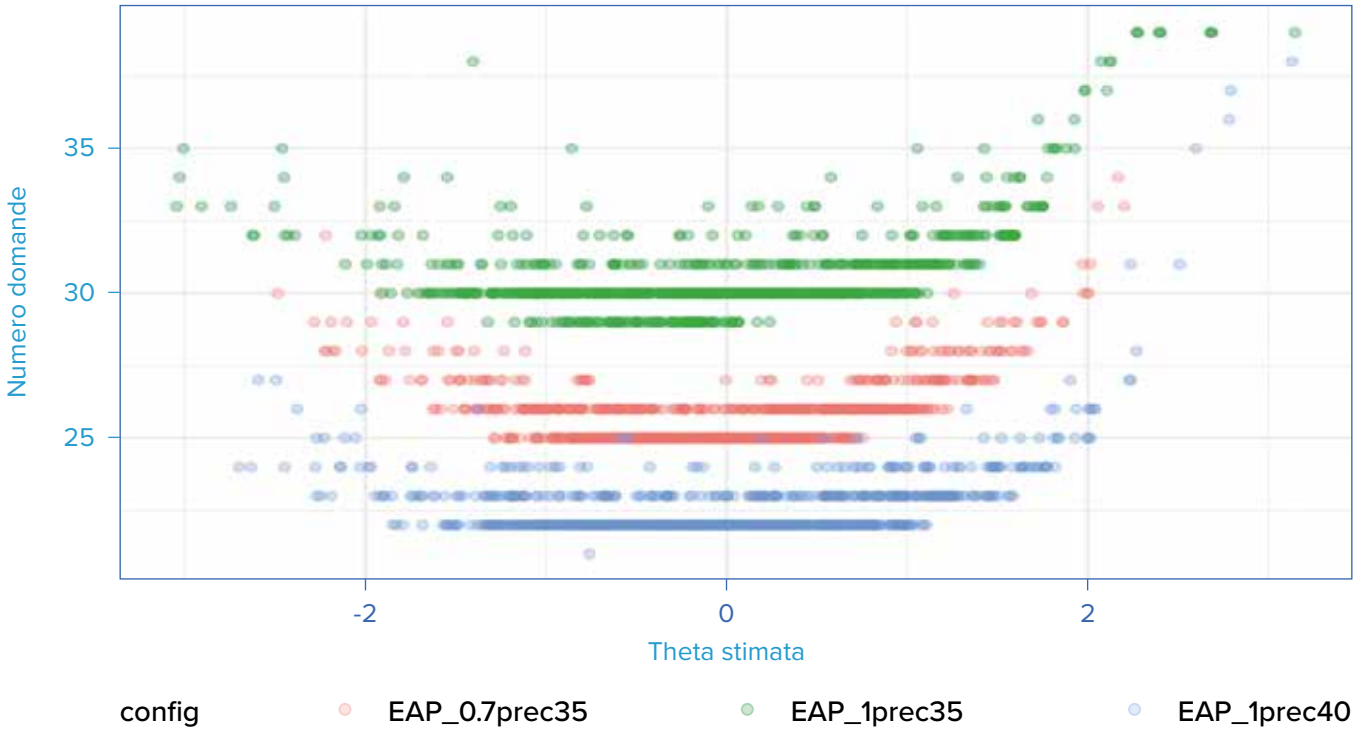
Confronto stimatori



Distribuzione numero item



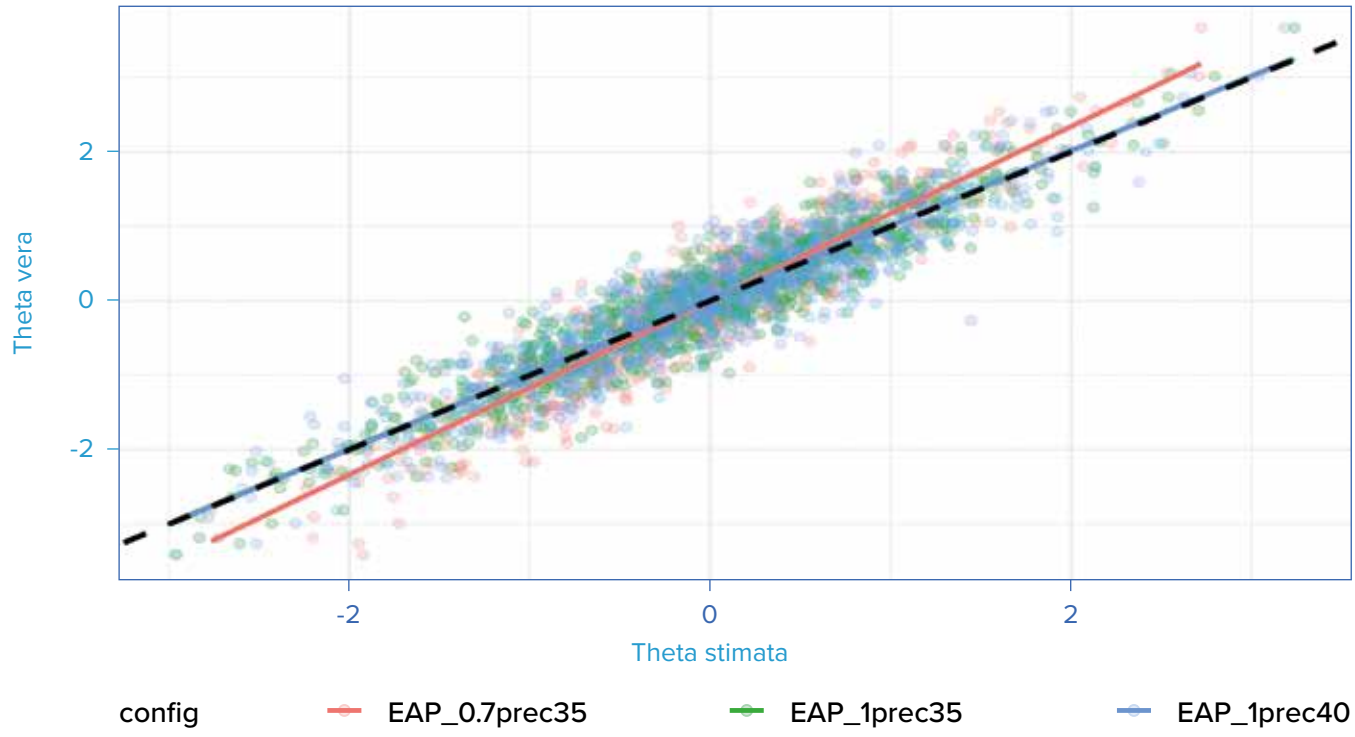
Numero domande vs Theta stimata



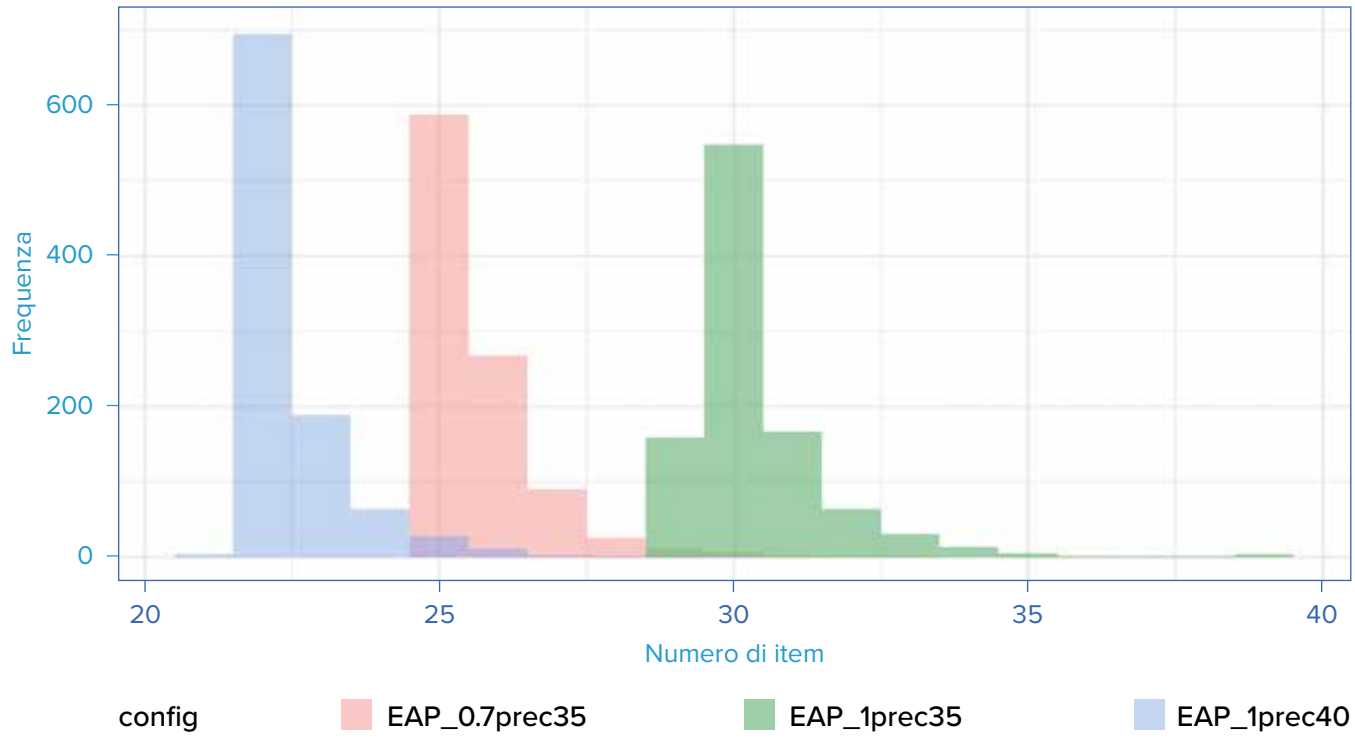
RATE-A	Correlazione	ES medio	N. item medio	DS item
EAP_0.7prec35	.93	.348	25.8	1.34
EAP_1prec35	.942	.348	30.65	1.73
EAP_1prec40	.925	.396	22.69	1.55

Figura 5. RATE-N: simulazione delle risposte con diversi parametri di settaggio

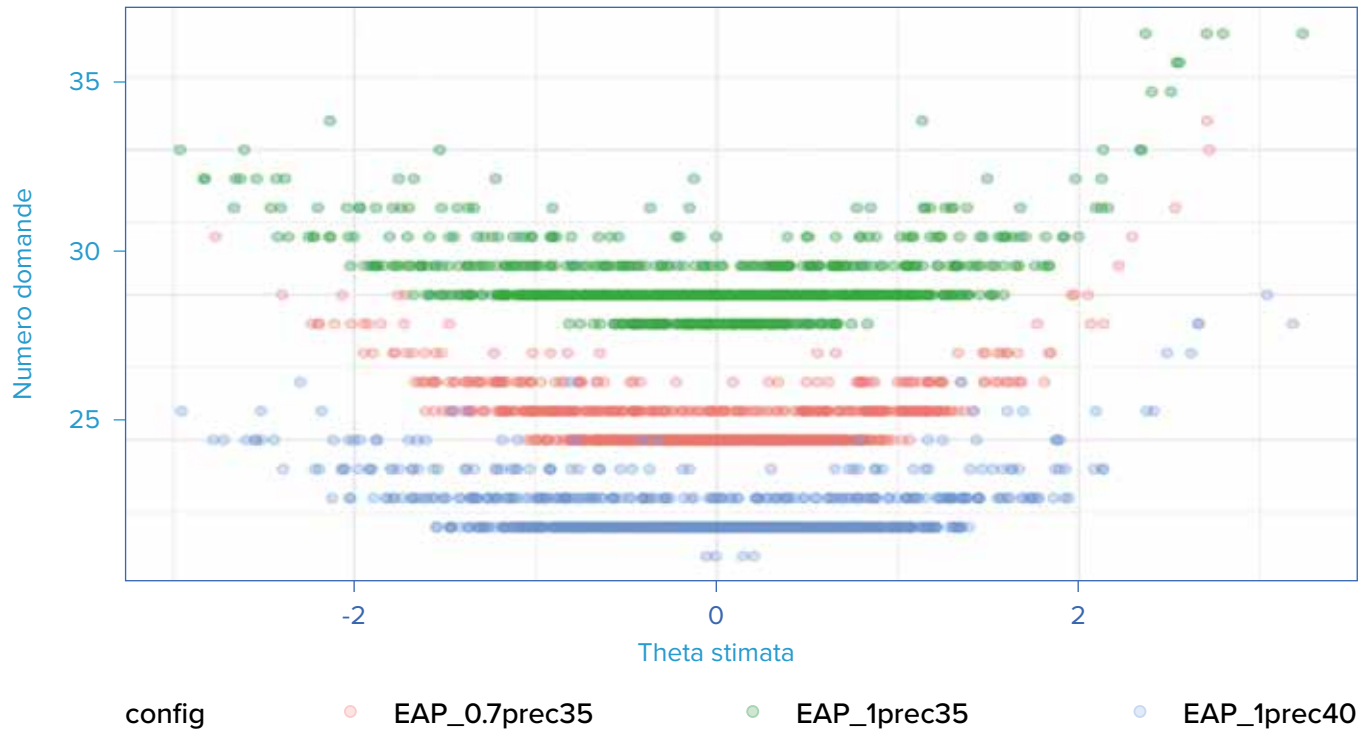
Confronto stimatori



Distribuzione numero item



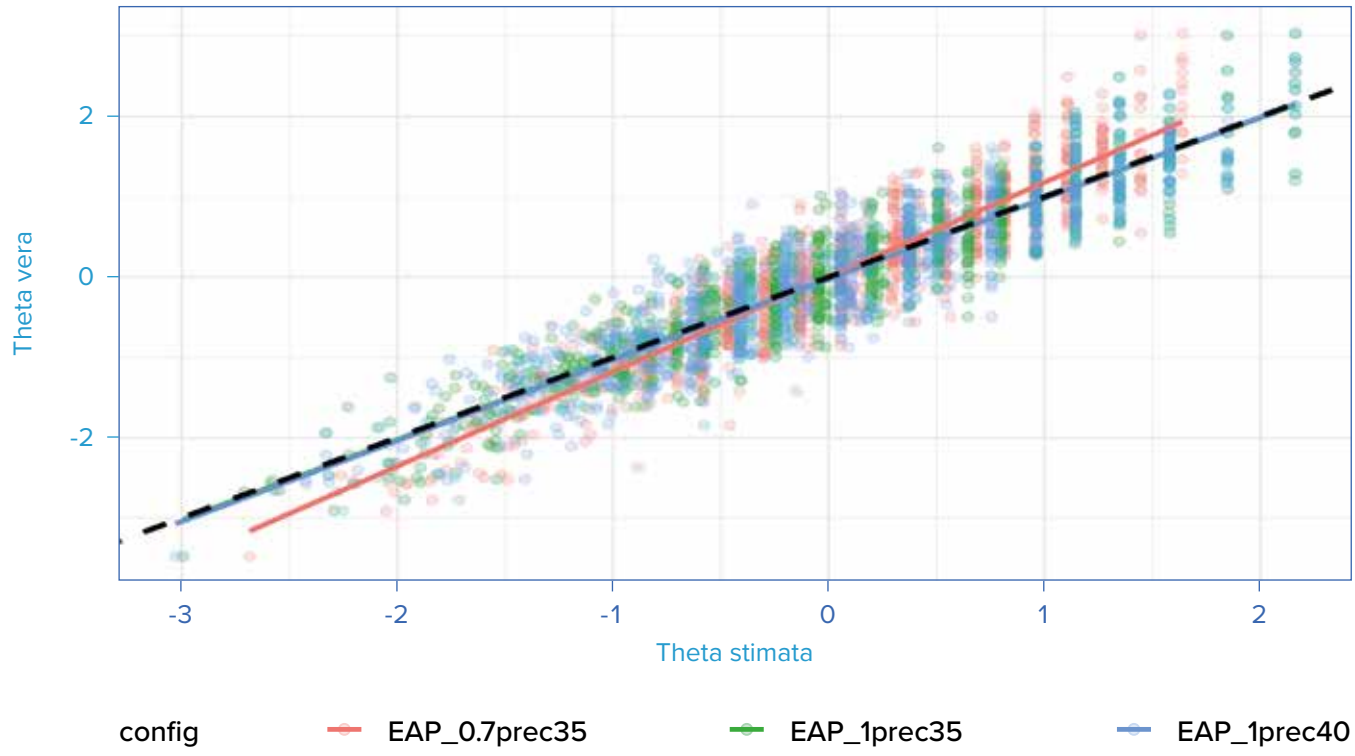
Numero domande vs Theta stimata



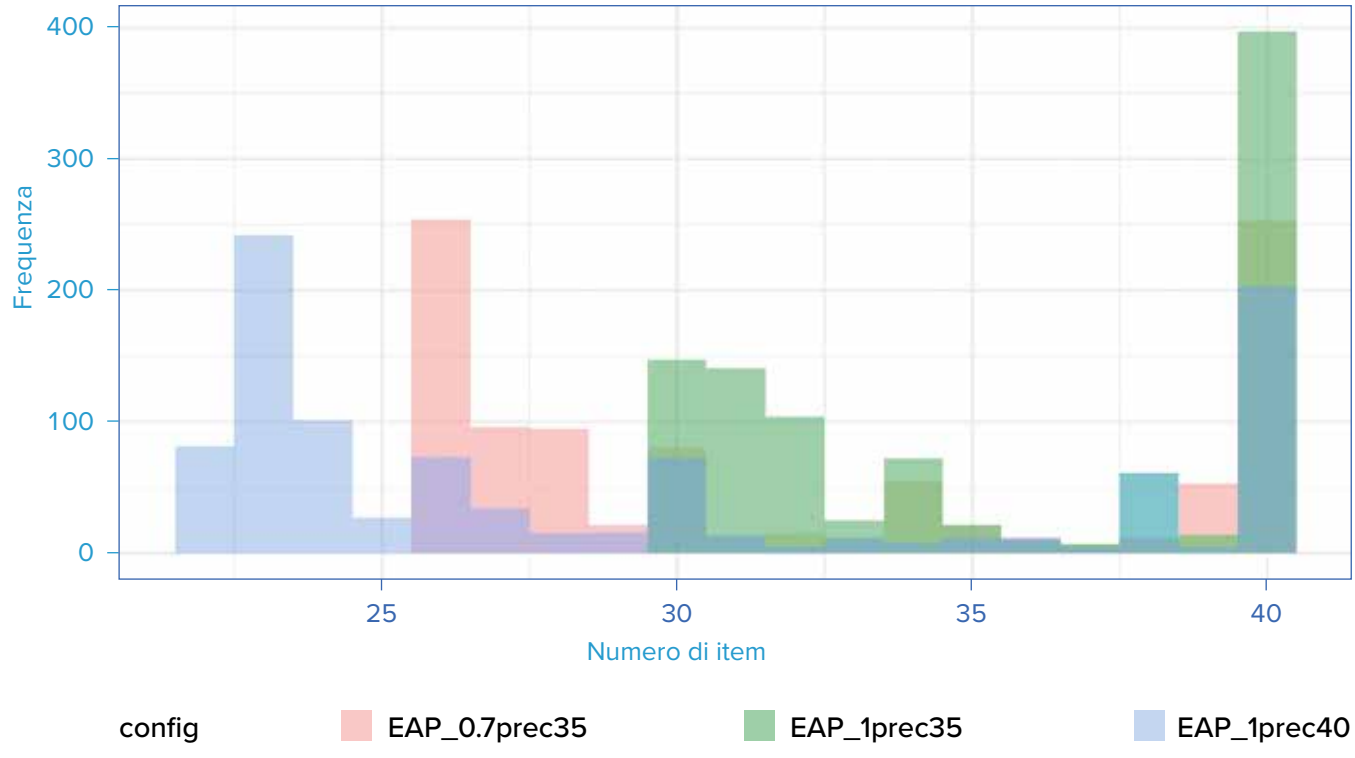
RATE-N	Correlazione	ES medio	N. item medio	DS item
EAP_07	.932	.348	25.72	1.14
EAP_1	.939	.347	30.43	1.42
EAP_1prec40	.92	.396	22.54	1.03

Figura 6. RATE-V: simulazione delle risposte con diversi parametri di settaggio

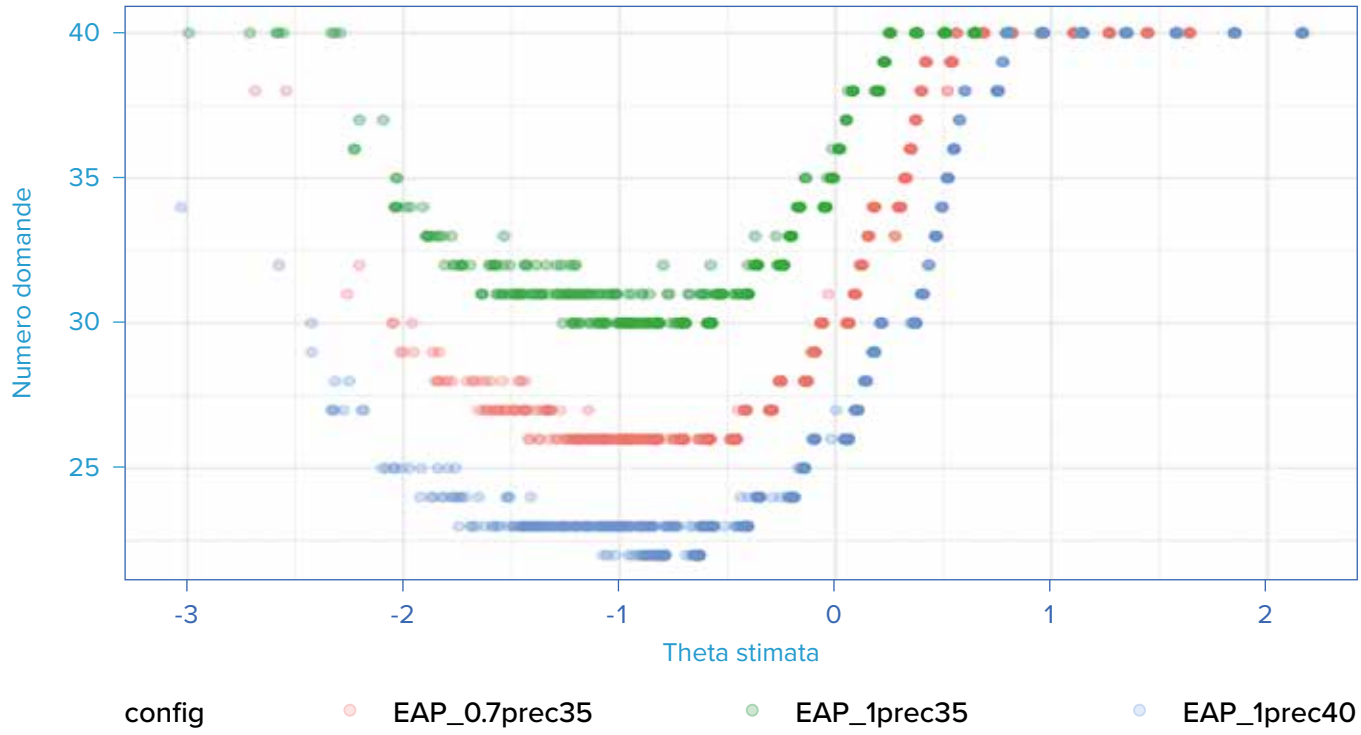
Confronto stimatori



Distribuzione numero item



Numero domande vs Theta stimata



RATE-V	Correlazione	ES medio	N. item medio	DS item
EAP_07	.917	.357	32.26	5.77
EAP_1	.925	.376	35.76	4.21
EAP_1prec40	.91	.409	29.42	6.86

RIFERIMENTI BIBLIOGRAFICI

Hartigan, J.A. e Wigdor, A.K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academy Press.

Linacre, J.M. (2002). What do infit and outfit, mean-squared and standardized mean? *Rasch Measurement Transactions, 16* (2), 878.

Schmidt, F.L. e Hunter, J.E. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86* (1), 162-173.

Schmidt, F.L., Hunter, J.E. e Outerbridge, A.N. (1986). The impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 73*, 46-57.

GIUNTI
PSYCHOMETRICS
www.giuntipsy.it
Bogotà | Bucarest | Budapest | Cairo | Campinas | Florence (HQ) | Istanbul | Jerusalem | Kyiv | Madrid | Mexico City | Milan
Moscow | Rio de Janeiro | Rome | San José de Costa Rica | San Sebastian | Santiago de Chile | São Paulo | Sofia | Turin