Università degli Studi di Milano-Bicocca
Corso di dottorato in Statistica - XXI ciclo

Tesi di Dottorato

La costruzione di una scala di misura per l'apprendimento della matematica nella scuola primaria

Giorgio Plazzi

30 novembre 2009

Indice

Introduzione

Il successo economico di ciascun individuo (nonché l'apporto individuale che questi può dare al progresso sociale) dipende largamente dalla provenienza e dalle modalità di produzione delle sue abilità e conoscenze. I genitori, l'influenza dei pari, le capacità individuali e la scolarizzazione sono solo alcuni dei fattori che contribuiscono allo sviluppo delle abilità e del capitale umano individuale e risulta molto complesso riuscire ad isolare il contributo di ciascuna componente nello sviluppo delle conoscenze e competenze individuali. In generale, gli indicatori sull'istruzione possono essere di tre tipi: indicatori di *input*, di processo o di *output*. Ciò che interessa, in un sistema educativo, sono gli *output* in vista degli *outcome* (Gori e Vittadini; 1999), ciò che gli individui sanno al termine del ciclo di studi intrapreso e come essi possono utilizzare la conoscenza acquisita.

Il Consiglio d'Europa, nel nuovo *framework* strategico per la cooperazione a livello europeo nel campo dell'istruzione e della formazione[1] ha confermato il ruolo centrale dell'istruzione e della formazione nell'agenda europea per l'occupazione e la crescita e ha identificato le sfide in cui tutti i sistemi di istruzione e formazione continentali saranno impegnati nei prossimi anni. Vincere queste sfide significa rendere l'apprendimento permanente e la mobilità degli studenti una realtà, migliorare la qualità e l'efficienza dell'istruzione e della formazione, promuovere l'equità, la coesione sociale e la cittadinanza attiva, sostenere la creatività e l'innovazione, anche imprenditoriale, ad ogni livello dell'istruzione e della formazione. Inoltre, la comunicazione "Efficienza e equità nei sistemi europei di istruzione e formazione" adottata 1'8 settembre 2006 dalla Commis-

[1]ET 2020, http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/educ/107622.pdf

sione europea, ha affermato una serie di principi significativi per lo sviluppo del servizio di istruzione e formazione.

Innanzi tutto, la Comunicazione afferma che "la combinazione tra autonomia locale degli istituti e sistemi di responsabilizzazione centralizzati può migliorare il rendimento degli studenti. Occorre tuttavia progettare i sistemi di responsabilizzazione in modo da garantire il pieno impegno a favore dell'equità e da evitare le conseguenze locali potenzialmente inique di decisioni decentrate, ad esempio relative alla delimitazione delle zone di utenza scolastica". Inoltre, il medesimo documento sottolinea come "l'accesso gratuito all'istruzione terziaria non garantisce necessariamente l'equità. Per rafforzare sia l'efficienza che l'equità gli Stati membri dovrebbero predisporre condizioni adeguate e incentivi per promuovere investimenti più elevati da parte delle fonti pubbliche e private, comprese, se del caso, anche le tasse di frequenza associate a misure finanziarie di accompagnamento per i soggetti svantaggiati." La Commissione, infine, richiama gli Stati membri a sviluppare una "cultura della valutazione" per capire a fondo e controllare ciò che avviene nei diversi sistemi (Comunicazione Della Commissione al Consiglio e al Parlamento Europeo. Efficienza ed equità nei sistemi europei di istruzione e formazione, 8-9-2006).

Si profila pertanto un tentativo di migliorare la qualità dei sistemi costituendo dei "quasi mercati" dell'istruzione rispettosi del principio di sussidiarietà[2]. La capacità di affrontare adeguatamente quest'impegno progettuale, tuttavia, dipende anche da una corretta concettualizzazione dell'oggetto del servizio, l'istruzione, e dalla corretta strutturazione del problema, ovvero co-

[2]Il principio di sussidiarietà entra prepotentemente sulla scena del diritto europeo con il Trattato di Maastricht in cui, all'art. 3b, viene sancito tra i principi cardine dell'Unione Europea che "la Comunità interviene secondo il principio di sussidiarietà, soltanto se e nella misura in cui gli obiettivi dell'azione prevista non possono essere sufficientemente realizzati dagli Stati membri e possono dunque, a motivo delle dimensioni e degli effetti dell'azione in questione, essere realizzati a livello comunitario". Le funzioni pubbliche, secondo il principio di sussidiarietà, spettano di regola ai soggetti che sono più vicini alla popolazione, e quindi ai bisogni ed alle risorse, e solo in via di eccezione possono essere in capo a soggetti collocati in posizioni via via più distanti dalla comunità locale. Un concetto che implica due livelli di lettura: quello della sussidiarietà verticale (fra istituzioni pubbliche) e quello della sussidiarietà orizzontale (fra istituzioni pubbliche e società civile, organizzata nella formazioni sociali).

me migliorare la qualità del sistema per migliorare la qualità dei risultati degli studenti. Centrale, in questo senso, è la capacità dei sistemi di produrre misure affidabili su quali siano effettivamente i risultati degli studenti in modo da poter informare le *policy*. In termini di sistema, le misure di *output* risultano essere gli indicatori maggiormente informativi per sviluppare sistemi di *evidence-based policy* e i paesi si stanno pertanto - e non senza contraddizioni - muovendo verso la costruzione di sistemi di *accountability* basati primariamente su tre aspetti[3]:

1. miglioramento delle conoscenze e competenze degli studenti misurate attraverso test standardizzati;

2. riduzione degli abbandoni scolastici precoci;

3. miglioramento dei tassi di raggiungimento del diploma di scuola secondaria superiore.

Negli ultimi anni si sta dedicando particolare enfasi al primo aspetto, anche grazie allo sprone dato dalle indagini internazionali, prima fra tutte l'indagine OCSE-PISA (*Programme for International Student Assessment*), promossa dall'Organizzazione per la Cooperazione e lo Sviluppo Economico (OCSE) per accertare le competenze dei quindicenni scolarizzati. PISA ha l'obiettivo di verificare se e in che misura i giovani quindicenni scolarizzati abbiano acquisito alcune competenze giudicate essenziali per svolgere un ruolo consapevole e attivo nella società e per continuare ad apprendere per tutta la vita (*lifelong learning*). L'indagine accerta il possesso di competenze nelle aree della comprensione della lettura, della matematica e delle scienze. L'attenzione non si focalizza tanto sulla padronanza di determinati contenuti curricolari, ma piuttosto sulla misura in cui gli studenti sono in grado di utilizzare competenze acquisite durante gli anni di scuola per affrontare e risolvere problemi e compiti che si incontrano nella vita quotidiana e per continuare ad apprendere[4].

Il primo vantaggio portato da queste indagini internazionali è senz'altro stato quello di riuscire, per la prima volta, a rendere comparabili i risultati

[3]si veda per esempio: Ministero dell'Economia e delle Finanze, Ministero della Pubblica Istruzione, Settembre 2007, Quaderno Bianco sulla scuola, www.pubblicaistruzione.it

[4]INVALSI; http://www.invalsi.it/ric-int/Pisa2006/sito/index.htm.

degli studenti a livello internazionale introducendo così uno strumento importante per ragionare sulla qualità dei sistemi scolastici a livello mondiale e per riuscire a comprendere quali siano gli strumenti che si possono adottare per migliorarli. L'Italia è uno dei paesi che per primi ha aderito a queste indagini internazionali ed è anche uno dei paesi che partecipa al maggior numero di progetti (il citato OCSE-PISA, ma anche IEA-TIMSS sull'apprendimento della matematica e delle scienze, IEA-PIRLS sulla lettura, *IEA-ICCS* sulla cittadinanza attiva, solo per citare i maggiori). Tuttavia, l'Italia è anche uno dei paesi che, fino a tempi recenti, si è mostrato più impermeabile a recepire le evidenze portate dai risultati raggiunti in queste indagini. Ciò è stato dovuto in parte a una tendenziale estraneità della tradizione pedagogica nazionale dall'utilizzo degli strumenti di valutazione propugnati dalle citate indagini internazionali, in parte al fatto che, da queste indagini internazionali, il sistema scolastico italiano ne esce perdente e con risultati sensibilmente inferiori alla media mondiale.

Limitando la disamina ai risultati dell'indagine PISA 2006, i dati confermano la debolezza del sistema scolastico italiano in termini di comparazione internazionale e le forti differenze esistenti al suo interno, sia tra aree geografiche del paese sia tra tipi di scuole. In PISA i punteggi conseguiti dai singoli studenti sono "tradotti" in scale di competenza, a ogni gradino delle quali corrisponde un diverso e crescente grado di competenza. In PISA 2006, il punteggio medio degli studenti italiani nella scala complessiva di scienze è pari a 475 (con una deviazione standard pari a 96), contro una media OCSE pari a 500 (deviazione standard pari a 95); la media dei 25 paesi dell'Unione Europea partecipanti a PISA 2006 è pari a 497; la differenza tra il punteggio degli studenti maschi e il punteggio delle studentesse non è statisticamente significativa. In Italia, il 25,3% degli studenti si colloca al di sotto del livello 2 della scala complessiva di scienze, che è stato individuato come il livello al quale gli studenti dimostrano di possedere il livello base di competenza scientifica (media OCSE 23,2). Meno del 5% degli studenti si colloca nei due livelli più elevati della scala (media OCSE 8,8). Anche per la matematica e per la lettura, il livello medio di prestazione degli studenti del nostro paese è inferiore a quello della media OCSE (462 rispetto a 498 per la matematica; 469

rispetto a 492 per la lettura). Per quanto riguarda i livelli di competenza in matematica, il 32,8% degli studenti italiani si colloca al di sotto del livello 2 della scala, che è stato individuato in PISA come il livello al quale gli studenti dimostrano di possedere le competenze basilari di matematica (media OCSE 21,3); solo il 6,3% degli studenti si colloca nei due livelli più elevati della scala (media OCSE 13,3). Per quanto riguarda la lettura, il 50,9% dei nostri studenti si colloca al di sotto del livello 3 della scala complessiva di lettura, che è stato individuato in PISA come il livello al quale gli studenti dimostrano di possedere le competenze di lettura basilari necessarie per confrontarsi in modo efficace con contesti e situazioni di vita quotidiana che richiedono l'esercizio di tale competenza (media OCSE 42,8); il 5,2% degli studenti si colloca nel livello più elevato della scala (media OCSE 8,6).

In Italia il 52,1% della varianza totale è spiegata dalla varianza tra scuole. Questo significa che più della metà delle differenze di prestazione degli studenti si deve alle differenze esistenti tra le diverse scuole frequentate. La media OCSE rispetto a questo indicatore è invece più bassa, pari al 33,1%. I paesi all'interno dei quali si riscontra la percentuale più bassa (meno del 20%) della varianza totale attribuibile alle scuole sono l'Australia, il Canada, la Danimarca, la Finlandia, l'Islanda, l'Irlanda, la Nuova Zelanda, la Norvegia, la Polonia, la Spagna, la Svezia, il Regno Unito. In alcuni di questi paesi la prestazione media degli studenti è superiore alla media OCSE. Come sottolineato nel rapporto internazionale di PISA 2006, questo significa che in tali paesi è possibile fare affidamento su scuole relativamente omogenee, e, soprattutto, dimostra che assicurare agli studenti contesti di apprendimento il più possibile equivalenti fra di loro non è incompatibile con il raggiungimento di risultati qualitativamente elevati. La situazione è molto diversa nel nostro paese, dove, a una elevata varianza tra scuole, si associa un livello di prestazione degli studenti inferiore alla media OCSE. Questo starebbe a indicare che nessuno dei due obiettivi sopra indicati (qualità e omogeneità) sembra essere raggiunto, con tutte le conseguenze che ne derivano in termini non soltanto di efficienza, ma anche - e forse soprattutto - di (non) equità. Questa valutazione è ulteriormente rafforzata dal confronto tra macroaree geografiche e tra tipi di scuole, che consente di rilevare le seguenti differenze principali: gli studenti di

liceo (media=518) conseguono risultati molto migliori d quelli di tutti gli altri indirizzi di studio, e migliori anche della media OCSE (500), seguiti dagli studenti degli istituti tecnici (media=475) e da quelli degli istituti professionali; il punteggio medio conseguito dagli studenti varia dal Nord al Sud del paese. Nell'ordine: Nord Est 520, Nord Ovest 501, Centro 486, Sud 448, Sud Isole 432; al di sopra della media OCSE si collocano gli studenti dei licei del Nord Ovest, del Nord Est e del Centro; gli studenti degli istituti tecnici del Nord Ovest e del Nord Est.

Sebbene per diversi anni queste evidenze non si siano tradotte in altrettanto fondamentali decisioni di *policy*, l'emergenza educativa che insiste sul Paese fa sì che vi sia ormai una crescente consapevolezza sulla necessità di intervenire con decisione per migliorare i risultati dell'Italia nel panorama internazionale. In questo senso si possono leggere le scelte operate con i PON[5] rivolti alle quattro regioni italiane dell'Obiettivo Convergenza (Calabria, Campania, Pu-

[5]Il PON (Programma Operativo Nazionale "La Scuola per lo Sviluppo") è uno dei 7 programmi operativi nazionali previsti dal Quadro Comunitario di Sostegno (QCS) finanziati dai Fondi Strutturali Obiettivo 1. La titolarità di questo programma è del Ministero dell'istruzione - Direzione Generale per gli Affari Internazionali dell'Istruzione Scolastica, Ufficio V - che funge da Autorità di Gestione. Il PON Scuola si avvale di due Fondi, il Fondo Sociale Europeo (FSE) e il Fondo Europeo di Sviluppo Regionale (FESR), e ha come ambito di riferimento territoriale le scuole pubbliche di 6 Regioni del Mezzogiorno, ossia: Basilicata, Calabria, Campania, Puglia, Sardegna e Sicilia. Il piano finanziario, inizialmente previsto per un importo pari a 718,557 MEURO, ha beneficiato di un incremento di 111 Meuro a decorrere dal 2004, a seguito del raggiungimento di tutti gli indicatori previsti per la riserva di premialità comunitaria e nazionale, e ammonta oggi, complessivamente, a 830.Meuro. Si tratta dunque di un vasto piano di sostegno finanziario allo sviluppo del sistema di istruzione e formazione delle regioni del mezzogiorno che ha come obiettivi di grande rilievo:

- la riduzione del fenomeno della dispersione scolastica;
- lo sviluppo della società della conoscenza e dell'informazione;
- l'ampliamento delle competenze di base;
- il sostegno alla mobilità dei giovani e lo sviluppo degli strumenti per garantirla;
- l'integrazione con il mondo del lavoro (stage, accreditamento competenze, certificazione);
- lo sviluppo dell'istruzione permanente e lo sviluppo di una cultura ambientale;
- la formazione dei docenti e del personale scolastico;
- il rafforzamento delle pari opportunità di genere.

glia e Sicilia), che mirano a migliorare la qualità del servizio scolastico elevando il livello di competenza degli studenti in lettura e matematica, basandosi sui dati dell'indagine PISA che mostra come in Italia - e in particolare in alcune aree del mezzogiorno - sono pochi i quindicenni scolarizzati che raggiungono i livelli di eccellenza, mentre troppi di loro si fermano ai livelli più bassi delle scale di competenza. Analogo, inoltre, è il rafforzamento del cosiddetto SNV (Servizio Nazionale di Valutazione) gestito dall'INVALSI, l'Istituto Nazionale di Valutazione del Sistema di Istruzione e formazione[6]. Il sistema SNV nasce da un percorso iniziato nel 2001 con i cosiddetti Progetti Pilota che - come emerge dalla relazione al Ministro del prof. Giacomo Elias, allora Presidente dell'INVALSI - si proponevano di simulare sul campo l'applicazione del modello di Servizio Nazionale di Valutazione dell'Istruzione, al fine di verificarne la praticabilità organizzativa e di consentirne la stima dei costi. Il modello doveva:

- misurare, scuola per scuola, il grado di raggiungimento degli obiettivi nazionali stabiliti dall'Alta Direzione (Ministro), integrando gli obiettivi nazionali con quelli dell'autonomia;

- consentire d'individuare tempestivamente e sistematicamente (annualmente), scuola per scuola, gli eventuali scostamenti rispetto agli obiettivi definiti al punto precedente e di intervenire, ai diversi livelli di responsabilità, con le necessarie azioni correttive e allocazioni di risorse, al fine di ottenere il miglioramento continuo del sistema dell'istruzione nazionale;

- utilizzare parametri coerenti con quelli usati dai servizi di valutazione comunitari e internazionali;

- richiedere risorse congruenti con le esigenze di bilancio.

[6]Con la direttiva n.75 del 15.9.2008 il Ministro dell'Istruzione, Università e Ricerca ha chiesto all'INVALSI di "provvedere [...] alla valutazione degli apprendimenti tenendo conto delle soluzioni e degli strumenti adottati per rilevare il valore aggiunto da ogni singola scuola in termini di accrescimento dei livelli di apprendimento degli alunni". La direttiva n.75 prevede, inoltre, nella prospettiva indicata dalla direttiva triennale n. 74, che per il presente anno scolastico, la rilevazione avvenga nel II e nel V anno della scuola primaria, per essere poi estesa, gradualmente ed entro il 2011, a tutti gli altri ordini di scuola.

La creazione dei Progetti Pilota e l'introduzione, alcuni anni or sono, in Italia di una prospettiva di valutazione esterna degli apprendimenti ebbe effetti variabili. In sé fu un fatto positivo che, nonostante le tensioni, permise di aprire lentamente la porta all'ingresso anche nel nostro paese di una cultura della valutazione. Tuttavia, la novità del sistema, unitamente ai risultati negativi che emergevano dalle citate indagini internazionali e il dubbio sull'uso che si sarebbe fatto delle informazioni raccolte (si temeva infatti che, nonostante le assicurazioni in proposito, queste venissero utilizzate per valutare l'operato degli insegnanti), creò un clima di diffidenza che costrinse a numerosi compromessi per quanto riguarda le caratteristiche dello strumento. Infatti, per assicurare gli operatori sul fatto che le prove erano costruite per scopi informativi e non sarebbero state utilizzate con finalità di controllo, il Gruppo di Lavoro INVALSI decise di:

1. svolgere i test a inizio anno al fine di informare gli insegnanti sul livello degli studenti in ingresso cosicchè i docenti potessero poi definire la didattica da implementare,

2. utilizzare test non confrontabili tra loro nel tempo in cui l'unica informazione a disposizione della scuola è di tipo spaziale (ovvero posizione nel *ranking*, cioè la posizione che ogni scuola ricopre rispetto a tutte le altre) e ciò che la scuola può vedere nel tempo è solo la sua variazione di posizione nel *ranking* ma non si possono fare ulteriori inferenze.

Questa tesi nasce proprio dall'interesse metodologico generato dalla necessità di risolvere uno di questi compromessi, ovvero la non confrontabilità nel tempo dei risultati degli studenti. L'utilizzo da parte dell'INVALSI di test con item differenti nelle diverse edizioni è stato senz'altro motivato dalla volontà di ridurre il rischio di "obsolescenza"[7] delle prove utilizzate nei Progetti Pilota e nelle diverse edizioni del Servizio Nazionale di Valutazione. Questo comporta che se le prove dell'anno t e quelle dell'anno t+1, somministrate a studenti

[7]Si dice che un test è "obsoleto" se il suo uso ripetuto aumenta la probabilità di risposta esatta ai diversi item, senza che questo implichi un reale aumento dei livelli di competenza latenti. In sostanza: distribuendo lo stesso test più volte gli studenti vengono a conoscenza delle risposte esatte.

dello stesso livello scolastico, hanno difficoltà differenti (cosa che non può essere esclusa a priori), le percentuali di risposte esatte, nei due anni, dipendono dalla diversa difficoltà, oltre che dalle differenze nei livelli di competenze, vero oggetto del monitoraggio. Anche se i confronti spaziali possono risultare ancora validi, una tale eventualità (differenze nella difficoltà dei test) rende estremamente problematico l'uso dei dati raccolti a fini di analisi diacroniche volte, ad esempio, a valutare se i livelli delle competenze degli studenti stiano migliorando o peggiorando. In quest'ottica anche il sistema di valutazione delle scuole che l'INVALSI ha adottato fino ad oggi, basato sul *ranking*, viene inficiato; esso non tiene conto né dell'effettivo livello di apprendimento raggiunto dagli studenti, né del valore aggiunto apportato dagli insegnanti.

Dall'edizione 2008-2009, il Servizio Nazionale di Valutazione ha preso nota del problema e d'ora in avanti dovrebbe essere possibile effettuare anche confronti temporali. Tuttavia, permane il problema di capire se e come possono essere utilizzati i patrimoni informativi prodotti nell'arco dei precedenti 7 anni. Ad ulteriore complicazione della questione, le suddette basi dati sono anche affette da problemi di *validity* e *reliability* in quanto, da una parte, non si è conservata memoria di come siano stati costruiti e validati gli item inseriti nei questionari e, dall'altra, la diffidenza - se non aperta contestazione da parte del mondo scuola - rispetto all'operato INVALSI ha fatto sì che diverse prove venissero quasi boicottate rendendo i risultati prodotti scarsamente affidabili[8].

In questo elaborato si presentano i risultati di un progetto sviluppato per proporre una soluzione a questi problemi, soluzione costituita dalla creazione di una scala di misura degli apprendimenti che goda delle proprietà di *oggettività specifica*, *additività* e *uniformità dell'unità di misura* su tutta la scala mediante la "equalizzazione" degli item nei test dei diversi periodi e dei diversi livelli, sfruttando le tecniche basate sull'analisi di Rasch (l'unico modello IRT in grado di soddisfare il principio dell'*oggettività specifica*). L'equalizzazione consiste nello stimare la difficoltà relativa di tutti gli item somministrati nelle varie prove, in modo da poterli collocare sulla stessa scala e valutare così anche il grado di difficoltà medio delle prove dei Progetti Pilota e SNV. Una volta

[8]Si veda in questo senso: Pasqualino Montanaro, La qualità dell'istruzione italiana, Banca d'Italia. http://www.aiel.it/bacheca/NAPOLI/D/montanaro.pdf

costruita una scala di misurazione, risulta immediato monitorare il livello di apprendimento raggiunto da ogni studente durante il suo percorso scolastico, rielaborare i dati raccolti nell'ambito dei Progetti Pilota e SNV utilizzando le stime dell'analisi di Rasch, qualora si evidenziasse una differenza statisticamente significativa nei test, e produrre misure dei livelli di competenze confrontabili anche nel tempo, oltre che nello spazio.

Nel progetto sono state analizzate le banche dati INVALSI sviluppate tra il 2004 e il 2006, si sono utilizzati alcuni degli item del pre-test SNV 2008-09 e alcuni degli item TIMSS IV 2007 e si è creato un nuovo questionario avvalendosi della collaborazione di alcune scuole volontarie all'interno delle quali sono stati testati gli studenti di IV elementare.

Nello specifico, il progetto si è occupato di:

- trattare le banche dati esistenti in base al protocollo Falzetti Plazzi Vidoni (2009) in modo da ripulirle dagli eventuali problemi di *teacher cheating* e problemi di definizione degli item;
- verificare quali sono gli item dei test di matematica che soddisfano il modello di Rasch e non presentano DIF territoriali quali quelli evidenziati su alcuni degli item della base dati 2005-06 dall'analisi Falzetti Plazzi Vidoni (2009);
- utilizzando gli item disponibili - che indagano le conoscenze e competenze dei fanciulli di II elementare, IV elementare, V elementare e I media inferiore - definire un ampio pool di item validi da poter successivamente utilizzare per costruire il test di *link*;
- sottoporre il pool di item a esperti della materia a livello elementare che possano identificare gli item utili per costruire un test da proporre a studenti di IV elementare;
- utilizzando gli item individuati, creare il nuovo test di link;
- somministrare il test di link agli studenti alle scuole campione;
- procedere all'equalizzazione dei due test del SNV di IV elementare (anni scolastici 2004/2005 e 2005/2006);

- costruire una scala di misurazione per l'apprendimento della matematica con cui misurare gli effettivi livelli di competenza degli studenti.

Il progetto ha avuto il pregio di essere il primo reale tentativo di interrogare il patrimonio informativo esistente nelle basi dati INVALSI avvalendosi della costruzione di una nuova banca dati mediante indagine sul campo.

Anche a causa dell'estrema problematicità delle banche dati utilizzate (in alcune province oltre il 30% dei dati è risultato inutilizzabile in quanto possibilmente affetto da fenomeni di *cheating* e anche i dati rimanenti, sebbene più verosimili, non possono comunque essere considerati affidabili), l'adozione di una logica diacronica nell'analisi dei risultati non ha comunque permesso di giungere a conclusioni definitive relativamente al fenomeno - se non altro in controtendenza rispetto alle evidenze empiriche internazionali - secondo cui i risultati degli studenti delle scuole dell'Italia del sud sarebbero in media migliori di quelli degli studenti dell'Italia del nord. Tuttavia l'analisi dei risultati ha permesso di identificare le differenze sostanziali nei livelli di difficoltà dei test erogati nelle diverse edizioni e ciò consente di prefigurare la possibilità di trasmettere alle scuole un patrimonio informativo importante per permettere loro di sviluppare strategie di miglioramento tese a far progredire nel tempo i risultati degli studenti.

L'elaborato si suddivide in quattro capitoli. Nel primo si presentano le teorie di riferimento per l'analisi dei test; dopo un breve accenno alla Teoria Classica si passano a esaminare le proprietà del modello di Rasch, l'unico, in ambito delle misurazioni di variabili latenti, a godere dell'*oggettività specifica* e della *sufficienza* degli stimatori, introducendo il *Simple Logistic Model* utilizzato in tutto il prosieguo della ricerca. Viene fatto un breve cenno agli altri modelli della famiglia dell'*Item Response Theory* (*2PLM e 3PLM*) spiegando perché, nonostante la loro somiglianza formale con il modello di Rasch, non possono essere assimilati a questo dal punto di vista concettuale.

Nel secondo capitolo si analizzano in dettaglio i Modelli di Rasch, concentrando l'attenzione sui metodi di stima dei parametri degli item (Massima Verosimiglianza Congiunta, Massima Verosimiglianza Condizionata, Massima Verosimiglianza Marginale e Stima per Dati Appaiati), sulle generalizzazioni

del *Simple Logistic Model* nel caso di item con più di due categorie di risposta (*Rating Scale Model*, *Partial Credit Model* e *Extended Logistic Model*) e nel caso di un sistema di riferimento che prevede anche l'intervento dei giudici nella correzione delle prove e sulle due principali violazioni al modello di Rasch: il *Differntial Item Functionig* e la *Local Dependency*. Il capitolo si conclude con la descrizione dei principali test per valutare l'adattamento dei dati osservati al modello teorico.

Il terzo capitolo si concentra sull'analisi e la pulizia dei database dell'INVALSI. Dopo aver verificato le forti incongruenze dei risultati delle prove del SNV del 2004/2005 e del 2005/2006 con quanto emerso dalle indagini internazionali si definiscono cinque indicatori per rilevare i casi (a livello di classe e di scuola) più anomali. Durante il processo di pulizia si studia anche l'adattamento degli item al modello di Rasch, individuando i quesiti che violano palesemente i suoi requisiti e selezionando le domande, utili alla costruzione della scala di misura, che, al contrario, si presentano conformi agli assunti di *monotonicità*, *indipendenza locale* e *invarianza delle stime* per sottogruppi di scolari. Il confronto finale con i risultati del TIMSS mette in luce che la pulizia dei database, pur migliorando l'affidabilità dei dati, non sortisce gli effetti (per quanto ambiziosi) sperati.

Nel quarto, e ultimo capitolo, vengono descritti, innanzi tutto, i criteri seguiti per la costruzione e la validazione del questionario di aggancio delle prove del SNV; si fa un breve riferimento a quali sono state le linee guida delle commissioni di lavoro dell'INVALSI nella definizione e costruzione dei questionari per il monitoraggio delle competenze in matematica e si descrivono i passaggi seguiti per costruire il test di link; dopodiché si descrivono, in dettaglio, le analisi per valutare la bontà dei dati raccolti, soffermandosi a lungo sulle analisi relative all'adeguamento degli item al modello (analisi delle distribuzioni degli item e degli scolari, analisi dei *Fit Residual* e osservazione delle *Item Characteristic Curve*). Una volta identificato il set di item con un buon adattamento si passa alla costruzione della scala di misura vera e propria, legando assieme le tre prove di matematica (SNV 2004/2005, SNV 2005/2006 e test di link) e stimando con un'unica analisi i parametri di tutti gli item. Il capitolo si chiude con due esempi applicativi che utilizzano le misure determinate con la

suddetta scala: il primo è uno studio che valuta di quanto crescono, ipoteticamente, nel tempo le abilità degli scolari, il secondo è un'analisi che descrive, a livello regionale, l'effetto della scuola e l'effetto della classe sulla varianza degli apprendimenti.

Colgo l'occasione per ringraziare il prof. David Andrich della Western Australia University che mi ha dato modo di lavorare per alcuni mesi in un ambiente stimolante e creativo e apprezzare la profondità della teoria di Rasch e la sua utilità quando applicata alla valutazione degli apprendimenti, la prof.ssa Maria Pia Perelli D'Argenzio per l'aiuto nella costruzione del test di *link*, la dr.ssa Ida Marais per il suo costante e prezioso supporto e la sua pazienza, il dr. Barry Sheridan e il dr. Steve Humpry per le loro indicazioni metodologiche, la dr.ssa Michela Battauz e il dr. Daniele Vidoni per la loro disponibilità e i loro consigli.

Capitolo 1

La misurazione di variabili latenti: teorie e metodi

1.1 La Teoria Classica dei Test (CTT)

La *Teoria Classica dei Test* (*Classical Test Theory*) nasce alla fine dell'Ottocento (Alfred Binet e altri, 1894) con l'intento di studiare l'affidabilità e la validità dei metodi fondati sui risultati dei questionari (*test scores*) come strumento di misura per ricavare soddisfacenti "misure" di caratteristiche psico-sociali, non direttamente osservabili in natura, delle persone esaminate. L'impiego su vasta scala e lo sviluppo della *CTT* ha inizio negli anni Trenta, anche se la formalizzazione dell'equazione fondamentale su cui tale teoria si basa viene proposta da Spearman (Spearman 1904a, 1904b) qualche decennio prima; tale equazione ipotizza una relazione lineare e additiva tra il punteggio osservato di un test (X), la misura della variabile latente (θ) e la componente casuale dell'errore (ϵ). L'idea deriva direttamente dal problema della misurazione nell'ambito delle scienze fisiche: infatti come osserva Taylor (Taylor 1982) "Nessuna quantità fisica (una lunghezza, un tempo, una temperatura, ecc.) può essere misurata con assoluta precisione. Operando con cura, possiamo essere capaci di ridurre le incertezze fisiche finché esse sono estremamente piccole, ma eliminarle del tutto è impossibile".

Proprio perché sempre affetta da errore, la "misura" di una variabile latente

necessita di tecniche che ne valutino l'attendibilità, che ne determinino il livello di confidenza attraverso una stima della correlazione delle "misure" dello stesso oggetto ottenute con due strumenti differenti (per esempio due test diversi) e per mezzo di una misurazione aggregata ricavata dalla media di più misurazioni della stessa variabile. A queste e a altre domande la *CCT* ha cercato di dare una risposta.

I limiti della *CCT* riguardano l'impossibilità

- di separare le caratteristiche delle persone da quelle degli item;
- di determinare, nella pratica, indici per la verifica dell'affidabilità dei test;
- di studiare il comportamento di un singolo individuo nei confronti di un singolo item in quanto si limita a fornire statistiche a livello generale dei test.

Tali limiti hanno portato allo sviluppo di nuove teorie di misurazione delle variabili latenti, la più nota delle quali è l'*Item Response Theory* (*IRT* - a uno, a due e a tre parametri), di cui il Modello di Rasch è un caso particolare che gode di specifiche e apprezzabili proprietà.

1.1.1 Il modello e i suoi assiomi

La *CTT* si basa su un modello relativamente semplice in cui l'*observed score*, il *true score* e l'*errore casuale* sono legati da una relazione lineare. Indicati con θ_ν l'abilità latente (*true score*) da misurare dell'individuo ν, con $X_{\nu j}$ la variabile osservata (*observed score*) per l'individuo ν nella prova j e con $\epsilon_{\nu j}$ l'errore casuale di misurazione, il modello si rappresenta con

$$X_{\nu j} = \theta_\nu + \epsilon_{\nu j}. \tag{1.1}$$

Nella formulazione teorica del modello si possono distinguere due tipi di esperimenti aleatori: uno che considera l'unità di osservazione (l'individuo) come campionaria, l'altro che considera il punteggio, per un determinato individuo, come campionario. L'unione dei due esperimenti implica che lo stesso

true score può essere considerato come un valore di una variabile casuale e la variabile errore una variabile aleatoria la cui distribuzione è un mistura delle distribuzioni degli errori delle singole unità d'osservazione. In quest'ottica, data un'unità di osservazione u_ν, θ_ν è definito come valore atteso di $X_{\nu j}$, cioè

$$\theta_\nu = E(X_{\nu j}|U = u_\nu), \tag{1.2}$$

e l'errore è definito come $\epsilon_{\nu j} = X_{\nu j} - \theta_\nu$ con

$$E(\epsilon_{\nu j}|U = u_\nu) = 0 \quad \text{e}$$

$$\sigma^2(X_{\nu j}|U = u_\nu) = \sigma^2(\epsilon_{\nu j})$$

il che significa che per la persona ν l'errore della misurazione effettuata con lo strumento j è una variabile casuale distribuita secondo una funzione $f(\epsilon_{\nu j})$, di media zero e con varianza pari alla varianza degli *observed scores* $\sigma^2(X_{\nu j})$.

1.1.2 Misurazioni parallele e affidabilità

Un concetto molto importante nella *CTT* (e utile, come si vedrà tra breve, almeno dal punto di vista teorico, per la stima del livello di *affidabilità* (*reliability*) di due o più misurazioni della stessa variabile) è quello di *misurazioni parallele*.

Definizione.
$X_{\nu j}$ e $X'_{\nu j}$ sono *misurazioni parallele* se

1. $E(X_{\nu j}) = E(X'_{\nu j}) = \theta_\nu$
2. $\sigma^2(\epsilon_{\nu j}) = \sigma^2(\epsilon'_{\nu j})$

Quindi $X_{\nu j}$ e $X'_{\nu j}$ sono *misurazioni parallele* se hanno lo stesso *true score* (θ_ν) e la stessa varianza degli *observed scores* ($\sigma^2(X_{\nu j}) = \sigma^2(X'_{\nu j})$).
La *CTT* si basa su ipotesi molto forti. Data una popolazione P, si ipotizza che:
1) il valore atteso degli errori sia nullo

$$E(\epsilon_{\nu j}) = 0;$$

2) la covarianza (o la correlazione) tra *true score* ed *errore casuale* sia nulla

$$Cov(\theta_\nu, \epsilon_{\nu j}) = 0;$$

3) la covarianza (o correlazione) tra gli errori di misurazioni distinte sia nulla

$$Cov(\epsilon_j, \epsilon_{j'}) = 0.$$

Da questi tre assunti segue che:
1) nella popolazione, il valore medio degli *observed scores* è uguale al valore medio dei *true scores*

$$E(X) = E(\theta + \epsilon) = E(\theta) \quad \text{e} \tag{1.3}$$

2) la varianza degli *observed scores* è uguale alla somma delle varianze dei *true scores* e degli errori:

$$\sigma^2(X) = \sigma^2(\theta + \epsilon) = \sigma^2(\theta) + \sigma^2(\epsilon) \tag{1.4}$$

L'affidabilità può essere pertanto pensata come il grado di precisione con cui si misura θ, dove per precisione si intende la stabilità del punteggio osservato di un individuo in ripetute e identiche somministrazioni dello stesso test. Nella *CTT* l'affidabilità è definita come $\rho^2_{X\theta}$, quadrato del coefficiente di correlazione tra gli *observed scores* e i *true scores*: formalmente $\rho^2_{X\theta}$ è chiamato il *coefficiente di affidabilità* della misurazione.

Dati due test paralleli $X_{\nu j}$ e $X'_{\nu j}$ si ha che

$$Cov(X_{\nu j}, X'_{\nu j}) = E[(\theta + \epsilon_{\nu j})(\theta + \epsilon'_{\nu j})] - E[(\theta + \epsilon_{\nu j})]E[(\theta + \epsilon'_{\nu j})] = \sigma^2_\theta \tag{1.5}$$

Quindi:

$$\rho(X_{\nu j}, X'_{\nu j}) = \frac{\sigma^2_\theta}{\sigma(X_{\nu j})\sigma(X'_{\nu j})} = \frac{\sigma^2_\theta}{\sigma^2_X} \tag{1.6}$$

cioè

$$\rho_{XX'} = \rho(X_{\nu j}, X'_{\nu j}) = \rho^2_{X\theta} \tag{1.7}$$

Il quadrato della correlazione tra gli *observed scores* e i *true scores* è uguale alla correlazione tra gli *observed scores* di due misurazioni parallele. Il coefficiente $\rho^2_{X\theta}$ serve a definire il concetto di affidabilità (*reliability*) mentre il coefficiente $\rho^2_{XX'}$ serve a valutarla empiricamente.
Inoltre, poiché $\rho^2_{X\theta} = \frac{\sigma^2_\theta}{\sigma^2_X}$ è possibile dare un'interpretazione dell'affidabilità come misura dell'ammontare della variazione degli *observed scores* attribuibile alla variazione dei *true scores*.
Riscrivendo il coefficiente di affidabilità come

$$\rho^2_{X\theta} = 1 - \frac{\sigma^2_\epsilon}{\sigma^2_X} \tag{1.8}$$

si ricava che

$$\sigma_\epsilon = \sqrt{\sigma^2_X(1 - \rho^2_{X\theta})} \tag{1.9}$$

Da quest'ultima espressione si comprende la relazione inversa che lega l'affidabilità all'errore: in casi estremi se $\rho^2_{X\theta} = 1$ tutta la variazione degli *observed scores* è attribuibile ai *true scores*, mentre se $\rho^2_{X\theta} = 0$ tutta la variazione degli *observed scores* è attribuibile all'errore.

Come si stima in pratica l'affidabilità? Un modo grossolano (e molto impreciso) consiste nel somministrare allo stesso gruppo di persone lo stesso identico test in due differenti momenti e di calcolare il coefficiente di correlazione dei punteggi totali (*test-retest reliability*).

Se si dispone di forme parallele dello stesso test, l'affidabilità può essere calcolata mediante il coefficiente di correlazione dei punteggi totali dei due test (*parallel-forms reliability*), valendo l'uguaglianza $\rho_{XX'} = \rho^2_{X\theta}$. Anche questo metodo, come il metodo del *test-retest*, non è esente da errori .
Il metodo di stima più diffuso è quello conosciuto come *Cronbach's alpha* (*internal consistency reliability*) originariamente ricavato da Kuder e Richardson (1937) per item dicotomici e poi generalizzato da Cronbach (Cronbach, 1951) per item a risposte ordinali di qualunque tipo.
L'idea su cui si basa consiste nel fatto che ogni singolo item del test, se confrontato con tutti gli altri, può essere usato per stimarne l'affidabilità. Per un

test di K item dicotomici, l'affidabilità soddisfa la seguente disuguaglianza

$$\rho_{XX'} >= KR-20 = \frac{K}{K-1}[\frac{\sigma_X^2 - \sum_{i=1}^{K} p_i(1-p_i)}{\sigma_X^2}] \tag{1.10}$$

dove σ_X^2 è la varianza degli *observed scores* totali del test nella popolazione e p_i è la proporzione delle persone che hanno risposto correttamente all'item i. $KR-20$ è nota come la *Formula 20 di Kuder-Richardson*. Poiché $KR-20$ è un limite inferiore di $\rho_{XX'}$, quando $KR-20$ è alto allora l'affidabilità del test è alta, mentre valori bassi di $KR-20$ non implicano una scarsa affidabilità. Se poi si hanno N test paralleli $X_1, X_2, ..., X_N$ e si considera il test

$$Y = \sum_{i=1}^{N} X_i$$

segue che

$$\sigma_{\theta_Y}^2 = N^2\sigma_{\theta_X}^2 \quad \text{e}$$

$$\sigma_{\epsilon_Y}^2 = N\sigma_{\epsilon}^2.$$

Poiché gli errori sono tra loro incorrelati, la varianza degli stessi cresce meno velocemente della varianza dei *true scores*; quindi l'affidabilità del test Y risulta pari a

$$\rho_{YY'} = \frac{N^2\sigma_\theta^2}{N^2\sigma_\theta^2 + N\sigma_\epsilon^2} \tag{1.11}$$

che, dopo alcuni passaggi algebrici, diventa

$$\rho_{YY'} = \frac{N\rho_{XX'}}{1+(N-1)\rho_{XX'}}. \tag{1.12}$$

Quest'ultima espressione rappresenta la formula di Spearman-Brown (*Spearman-Brown prophecy formula*) che esprime l'effetto della lunghezza sull'affidabilità del test ed è ampiamente usata in pratica quando si vuole valutare l'affidabilità di un test replicato (nella stessa forma o con test paralleli). Essa mette in luce la relazione non lineare che lega l'affidabilità alla lunghezza del test. Bisogna però sottolineare che, se l'ampiezza di un test molto affidabile, è incrementata aggiungendo diversi item "poveri", l'affidabilità che ne deriva sarà, molto pro-

babilmente, più bassa di quella di partenza.
La 1.12 può essere riscritta, come segue, esplicitando N per determinare il numero di replicazioni necessarie a raggiungere un predeterminato grado di affidabilità

$$N = \frac{\rho_{YY'}(1-\rho_{XX'})}{\rho_{XX'}(1-\rho_{YY'})}.$$

Si consideri ora il test X composto da K parti

$$X = X_1 + X_2 + ... + X_K.$$

Il suo coefficiente di affidabilità è

$$\rho_{XX'} = \frac{\sigma_\theta^2}{\sigma_X^2} = \frac{\sum_{i=1}^K \sigma_{\theta_i}^2 + \sum_{i=1}^K \sum_{i\neq j}^K Cov(\theta_i, \theta_j)}{\sigma_X^2}. \tag{1.13}$$

Le covarianze tra i *true scores*, uguali alle covarianze tra gli *observed scores*, possono essere approssimate nel modo seguente:
poiché

$$(\sigma_{\theta_i} - \sigma_{\theta_j})^2 \geq 0$$

segue che

$$\sigma_{\theta_i}^2 + \sigma_{\theta_j}^2 \geq 2\sigma_{\theta_i}\sigma_{\theta_j}$$

e poiché

$$\sigma_{\theta_i}\sigma_{\theta_j} \geq Cov(\theta_i, \theta_j) \quad \text{(il coefficiente di correlazione è } \leq 1)$$

segue che

$$(K-1)\sum_{i=1}^K \sigma_{\theta_i}^2 = \sum_{i<j}^{K-1}\sum_{j=1}^K (\sigma_{\theta_i}^2 + \sigma_{\theta_j}^2) \geq \sum_{i=1}^K \sum_{j\neq i}^K Cov(\theta_i, \theta_j).$$

Da quest'ultima espressione si ottiene

$$\rho_{XX'} = \frac{\sum_{i=1}^K \sigma_{\theta_i^2} + \sum_{i=1}^K \sum_{j\neq i}^K Cov(\theta_i, \theta_j)}{\sigma_X^2} \geq \frac{\frac{K}{K-1}\sum_{i=1}^K \sum_{j\neq i}^K Cov(\theta_i, \theta_j)}{\sigma_X^2} =$$

$$= \frac{\frac{K}{K-1}\sum_{i=1}^{K}\sum_{j\neq i}^{K} Cov(X_i, X_j)}{\sigma_X^2} =$$

$$= (\frac{K}{K-1})\frac{\sigma_X^2 - \sum_{i=1}^{K}\sigma_{X_i}^2}{\sigma_X^2}. \tag{1.14}$$

L'ultima espressione è conosciuta come *coefficiente* α o *Cronbach's* α. Se la 1.14 è riscritta nella forma

$$\alpha = (\frac{K}{K-1})(1 - \frac{\sum_{i=1}^{K}\sigma_{X_i}^2}{\sigma_X^2}) \tag{1.15}$$

si vede che essa generalizza il $KR-20$ del caso di item dicotomici espresso dalla 1.10.

1.1.3 Errore di misura e stima del *true score*

True score e variabile errore sono i concetti cardine su cui si fonda la CTT ma, in pratica, è quasi sempre impossibile determinarli con precisione; ciò che invece risulta possibile e utile è stimare le loro varianze mediante esperimenti ripetuti su campioni della stessa popolazione. I seguenti risultati sono alla base dell'analisi dell'errore della CTT:

$$1. \quad Cov(X, \theta) = Cov[(\theta + \epsilon)\theta] = \sigma_\theta^2 \tag{1.16}$$

e poiché vale la 1.5, la varianza della variabile θ può essere calcolata mediante la covarianza di due test paralleli;

$$2. \quad Cov(X, \epsilon) = Cov[(\theta + \epsilon)\epsilon] = \sigma_\epsilon^2 \tag{1.17}$$

e poichè la varianza degli *observed scores* è pari alla somma della varianza dei *true scores* e della varianza degli errori, quest'ultima può essere calcolata come differenza

$$\sigma_\epsilon = \sigma_X^2 - Cov(X, X') \quad \text{con } X \neq X'. \tag{1.18}$$

Inoltre

$$\rho^2_{X\theta} = \frac{\sigma^2_{X\theta}}{\sigma^2_X \sigma^2_\theta} = \frac{\sigma^2_\theta}{\sigma^2_X} \quad \text{e} \tag{1.19}$$

$$\rho^2_{X\epsilon} = \frac{\sigma^2_{X\epsilon}}{\sigma^2_X \sigma^2_\epsilon} = \frac{\sigma^2_\epsilon}{\sigma^2_X} \tag{1.20}$$

e quindi segue che

$$\rho^2_{X\theta} + \rho^2_{X\epsilon} = 1. \tag{1.21}$$

Il quadrato del coefficiente di correlazione tra gli *observed scores* e i *true scores*, $\rho^2_{X\theta}$, riflette l'ammontare dell'errore contenuto nella misurazione; più grande è la correlazione tra i *true scores* e gli *observed scores*, più piccolo è l'errore nella misurazione X_{ij}. Infatti dalla relazione $X = \theta + \epsilon$ risulta

$$\sigma^2_\epsilon = \sigma^2_X - \sigma^2_\theta \tag{1.22}$$

che scritta nella forma

$$\sigma^2_\epsilon = \sigma^2_X(1 - \frac{\sigma^2_\theta}{\sigma^2_X}) = \sigma^2_X(1 - \rho^2_{X\theta}) \tag{1.23}$$

mette in evidenza la relazione che intercorre tra σ^2_ϵ, σ^2_X e $\rho^2_{X\theta}$. La radice quadrata positiva di σ^2_ϵ è lo *standard error* della misurazione nella *CTT*. Come anzidetto $\rho^2_{X\theta}$ è uguale a $\rho_{XX'}$ per cui l'errore della misurazione si può esprimere come

$$\sigma_\epsilon = \sigma_X \sqrt{1 - \rho_{XX'}}. \tag{1.24}$$

L'errore σ_ϵ viene pertanto stimato mediante la

$$\widehat{\sigma}^2_\epsilon = \widehat{\sigma}^2_X(1 - r_{XX'}) \tag{1.25}$$

dove

$$\widehat{\sigma}^2_X = \frac{\sum_{\nu=1}^{N}(X_{\nu j} - \overline{X})^2}{N - 1} \tag{1.26}$$

è uno stimatore della varianza degli *observed scores* (σ^2_X) in un campione di N individui e

$$r_{XX'}$$

uno stimatore di $\rho_{XX'}$.

Assumendo che gli errori di misurazione siano distribuiti approssimativamente come una variabile normale, è possibile costruire un intervallo di confidenza, ad esempio al 95%, per il *true score* θ_ν:

$$x_\nu - 1.96\widehat{\sigma_\epsilon} \leq \theta_\nu \leq x_\nu + 1.96\widehat{\sigma_\epsilon}. \tag{1.27}$$

L'inconveniente di questa formula è che essa assume l'errore di misurazione uguale per ogni θ mentre in realtà l'errore varia da persona a persona.

Un'altra applicazione molto importante riguarda la regressione dei *true scores* sugli *observed scores*. Se si suppone che nella popolazione P il legame tra *true scores* e *observed scores* sia di tipo lineare

$$\widehat{\theta} = ax + b \tag{1.28}$$

allora a e b possono essere presi in modo che la somma dei quadrati delle differenze tra i *true scores* e le loro stime sia minima. La formula che si ricava è la regressione dei *true scores* sugli *observed scores*:

$$\widehat{\theta} = \frac{\sigma_\theta \rho_{X\theta}}{\sigma_X}(x - \mu_X) + \mu_\theta. \tag{1.29}$$

che riscritta, applicando le 1.6 e 1.7 e valendo l'ipotesi che $\mu_\theta = \mu_X$, diventa:

$$\widehat{\theta} = \rho_{XX'}x + (1 - \rho_{XX'})\mu_X \tag{1.30}$$

con uno *standard error* della stima pari a

$$\sigma_E = \sigma_\theta\sqrt{1 - \rho^2_{X\theta}} = \sigma_X\sqrt{\frac{\sigma^2_\theta}{\sigma^2_X}(1 - \rho^2_{X\theta})} = \sigma_X\sqrt{\rho_{XX'}}\sqrt{1 - \rho_{XX'}} = \sigma_\epsilon\sqrt{\rho_{XX'}}. \tag{1.31}$$

La (1.30) è nota come *formula di regressione di Kelley* (Kelley, 1947). Essa afferma che, per misure molto affidabili, la miglior stima (in termini di minimi quadrati) del *true score* dell'individuo è data dal suo *observed score*; man mano che l'affidabilità diminuisce il termine x perde peso mentre cresce il peso di

μ_X, fino al caso limite in cui l'affidabilità è nulla ($\rho^2_{X\theta} = \rho_{XX'} = 0$) e allora il *true score* stimato di ogni individuo della popolazione P coincide con la media degli *observed scores* μ_X.
Sono state mosse critiche di un certo rilievo a questo metodo di stima. Innanzitutto lo *standard error* della stima presuppone una varianza dell'errore di misurazione costante; poi il legame tra *true scores* e *observed scores* potrebbe non essere lineare; ancora lo stimatore, in generale, non è corretto (il valore atteso della *formula di Kelley* è uguale a θ solo quando il *true score* coincide con la media della popolazione); infine, la formula di regressione non fornisce stime precise se viene applicata a campioni di piccole dimensioni.

L'indice di correlazione degli *observed scores* di due test arbitrari è dato da

$$\rho_{XY} = \rho_{\theta_X \theta_Y} \sqrt{\rho_{XX'} \rho_{YY'}} \tag{1.32}$$

dove $\rho_{XX'}$ e $\rho_{YY'}$ indicano i coefficienti di affidabilità per il test X e il test Y, rispettivamente.
Se uno o entrambi i coefficienti di affidabilità sono minori di 1 - come avviene nella pratica - allora $\sqrt{\rho_{XX'} \rho_{YY'}} < 1$ e $\rho_{XY} < \rho_{\theta_X \theta_Y}$. In questo caso la correlazione tra le misurazioni osservate tramite X e Y risulta attenuata a causa della non perfetta affidabilità. Noti i coefficienti di affidabilità di X e di Y, la correlazione tra i *true scores* dei due test è determinata a partire dal coefficiente di correlazione degli *observed scores* mediante

$$\rho_{\theta_X \theta_Y} = \rho_{XY} [\frac{1}{\sqrt{\rho_{XX'} \rho_{YY'}}}] \tag{1.33}$$

dove si vede che la correlazione tra gli *observed scores* dei due test viene moltiplicata per una quantità che è sempre ≥ 1 (il reciproco della radice quadrata del prodotto degli indici di affidabilità di ciascun test); pertanto la correlazione tra *true scores* risulta, in generale, maggiore della correlazione osservata tra gli *observed scores*.

1.2 La teoria di Rasch

1.2.1 L'oggettività specifica

La base di partenza per la costruzione di una teoria alternativa è costituita dalla ricerca del matematico danese Georg Rasch (1960) il quale, nel porsi il problema di individuare ciò che caratterizza la superiorità delle scienze naturali rispetto a quelle umane, giunse alla conclusione che il concetto di "scienza" è legato alla possibilità di sviluppare metodi per trasformare osservazioni in misure, secondo regole che soddisfano il principio della *oggettività specifica*. In termini intuitivi tale principio si riferisce al fatto che i metodi di misurazione delle scienze naturali consentono di misurare caratteristiche specifiche di un soggetto senza che il processo di misurazione risulti influenzato da caratteristiche del soggetto diverse da quella di interesse, da altri soggetti e da peculiarità dello strumento utilizzato a tale scopo. Rasch chiarisce ulteriormente il concetto di *oggettività specifica* osservando che ogni procedimento di misura scaturisce sempre da un "confronto" tra gli elementi dell'insieme A - cioè i soggetti - e gli elementi dell'insieme B - ossia le prove. Quando gli elementi di A entrano in contatto con gli elementi di B dalle coppie $(a,\ b)$ scaturiscono i risultati che costituiscono l'insieme R. In alcuni casi questi possono essere dicotomici (la risposta è giusta o sbagliata), in altri politomici (come il grado di soddisfazione o il giudizio espresso su una scala da 1 a 4), ma in altri ancora possono essere di tipo discreto, come quando si misura l'altezza di una persona con un metro che contiene solo l'indicazione dei centimetri o come quando si conta il numero di errori in una composizione scritta.

Se il contatto tra il soggetto e la prova produce un risultato ben determinato $r = r\ (a,\ b)$, detto reazione, si dice allora che $F = (A, B, R)$ costituisce un *sistema di riferimento* (bifattoriale) *di tipo deterministico*. In altre situazioni, che sono tipiche delle scienze umane, ma anche della fisica quantistica, la reazione è influenzata da errori e fattori casuali per cui r è una variabile aleatoria con una certa distribuzione $\Pr(R = r) = f(a,\ b)$, che dipende dal soggetto e dalla prova: in questo caso si parla allora di *sistemi di riferimento di tipo probabilistico*. Dunque, nel caso di *sistemi deterministici* il principio di *oggettività*

specifica, come spiega Rasch (1977), è legato al fatto che quando si confrontano le reazioni $r_1 = r(a_1, b)$ e $r_2 = r(a_2, b)$ di due soggetti a_1 e a_2, conseguenti al contatto con una medesima prova b, tale confronto $u(r_1, r_2)$ può dipendere dalla particolare prova b prescelta: ossia $u(r_1, r_2) = (r(a_1, b), r(a_2, b))$. L'insieme dei possibili risultati u del confronto costituisce l'insieme U che non necessariamente è costituito da numeri. Ad esempio, dal confronto tra $r_1 = r(a_1, b)$ e $r_2 = r(a_2, b)$ dove a_1 e a_2 sono soggetti e b un pezzo di piombo che viene posto sull'altro piatto della bilancia, può scaturire il risultato (di tipo qualitativo) $u = (a_1$ è più pesante di $a_2)$ per il motivo che il piatto della bilancia "pende" dalla parte del soggetto quando a_1 e b entrano in contatto, mentre "pende" dalla parte del piatto contenente il pezzo di piombo quando sono a_2 e b a entrare in contatto.

Rasch afferma che un *sistema di riferimento di tipo deterministico* è caratterizzato dalla proprietà della *oggettività specifica* se la funzione $u(r_1, r_2) = u(r(a_1, b), r(a_2, b)) = v(a_1, a_2)$ non dipende da b per qualsiasi coppia di soggetti e per qualsiasi prova. L'*oggettività* si riferisce appunto al fatto che il risultato del confronto tra due soggetti dell'insieme A è indipendente dalla scelta della prova b con cui i due soggetti entrano in contatto e da qualsiasi altro elemento dell'insieme A. Il concetto di *specificità* serve a puntualizzare il fatto che l'oggettività di questi confronti è ristretta al sistema di riferimento F. Il concetto di *oggettività specifica* viene esteso al caso di sistemi di riferimento multifattoriali in cui la reazione deriva dal contatto tra tre o più fattori: quindi non solo due fattori come nel caso del soggetto e della prova, ma, ad esempio, tre fattori: soggetti, prove e giudici.

Il risultato a cui giunge Rasch è che nel caso in cui ogni soggetto, ogni prova e ogni reazione sono caratterizzati da un parametro di tipo scalare, cioè da un numero reale θ_ν, δ_i e $x_{\nu i}$, allora $x_{\nu i} = r(a_\nu, b_i) = q(\theta_\nu, \delta_i)$, allo stesso tempo, la funzione q deve soddisfare la condizione $u(q(\theta_\nu, \delta), q(\theta_{\nu'}, \delta)) = v(\theta_\nu, \theta_{\nu'})$ al fine di assicurare la proprietà di oggettività specifica per il sistema di riferimento in questione. Sotto ipotesi di regolarità molto generali sulla funzione q è possibile stabilire le condizioni di esistenza ed unicità, nonché le caratteristiche necessarie e sufficienti che la funzione di reazione $q(\theta_\nu, \delta_i)$ deve possedere per garantire l'oggettività specifica per F. Tali condizioni e

caratteristiche sono sintetizzate nel seguente teorema.

Teorema.
Sia dato un sistema di riferimento bifattoriale F caratterizzato da parametri scalari θ e δ e da una reazione costituita da una funzione scalare $q(\theta_\nu, \delta_i)$ regolare, allora l'esistenza di tre funzioni strettamente monotòne:

$$\theta' = \phi(\theta), \delta' = \psi(\delta), x' = \chi(x)$$

in grado di trasformare $q(\theta, \delta)$ in una relazione puramente additiva

$$x' = \theta' + \delta'$$

è necessaria e sufficiente affinché il sistema di riferimento F sia caratterizzato dalla *oggettività specifica*. Inoltre, se tali funzioni esistono, esse sono uniche a meno di trasformazioni lineari, e il confronto delle reazioni di due *soggetti* $u(x_{\nu i}, x_{\nu' i}) = v(\theta_\nu, \theta_{\nu'})$, o di due prove, $u(x_{\nu i}, x_{\nu j}) = w(\delta_i, \delta_j)$, si riduce semplicemente alle differenze:

$$x'_{\nu i} - x'_{\nu' i} = \theta'_\nu - \theta'_{\nu'} \quad \text{(per i soggetti)}$$

$$x'_{ni} - x'_{nj} = \delta'_i - \delta'_j \quad \text{(per le prove)}$$

Il teorema può essere esteso al caso di sistemi di riferimento con più di due fattori.

Se questa condizione è soddisfatta il sistema di riferimento F si dice caratterizzato da *additività latente* e, come Rasch osserva, questo è tipico di tutti i sistemi di misura a scala di intervallo (come ad esempio, la temperatura, in cui l'origine 0 è convenzionale, ma x gradi in più o in meno hanno lo stesso significato a qualsiasi livello della scala). Rasch riporta una serie di esempi in cui mostra come tutte le misure fondamentali della fisica soddisfino il principio dell'*oggettività specifica*.

L'intuizione di Rasch, relativa al fatto che la superiorità delle scienze naturali rispetto a quelle umane risieda nell'adozione di sistemi di misura carat-

terizzati dall'*oggettività specifica*, è del tutto giustificata se si considera che il metodo galileiano, sul quale le scienze naturali si fondano, prevede tra i suoi momenti essenziali la misurazione. Tali momenti sono:

a) l'**Ipotesi**, ossia la formulazione della teoria, applicabile a tutti i fenomeni dello stesso genere e non solo a quelli oggetto della sperimentazione;

b) l'**Esperimento**, che può essere dato sia dall'osservazione di fenomeni spontanei in natura sia dalla stimolazione adeguata della natura con mezzi appropriati. In tutti e due i casi l'essenza (e la novità) dell'approccio galileiano consiste nel ridurre l'esperimento a misurazione dei fenomeni. Per Galileo c'è una "ferma e costante" relazione tra i fenomeni, che è data dalla relazione tra causa ed effetto. Tuttavia, quello che interessa di tale relazione non è stabilire se tra essi esista una connessione nascosta, ma formulare matematicamente tale rapporto;

c) la **Verifica**. Affinché l'ipotesi risulti "vera" (cioè non falsificata secondo Popper (1959)), deve ritornare di nuovo alla pratica: deve cioè essere sperimentata, applicandola a casi sempre nuovi. Ciò comporterà un duplice vantaggio: renderà "certa" (fino a una eventuale falsificazione) l'ipotesi stessa e farà in modo che la scienza diventi tecnica, cioè che i principi teorici vengano applicati in strumenti che dapprima serviranno solo a verificare la teoria, ma che poi saranno applicati a tutti gli usi immaginabili.

Il maggior contributo di Rasch - la cui notorietà è legata all'omonimo modello per la trattazione dei test in ambito psicometrico - consiste nell'avere indicato al mondo delle scienze umane la via per elevarsi al rango delle scienze naturali. Costruendo teorie che, ove coinvolgano entità latenti per le quali non sia stato ancora individuato un adeguato strumento di misura, utilizzino solo misure che soddisfano il principio della *oggettività specifica*.

1.2.2 Oggettività specifica e sufficienza

In un sistema di riferimento probabilistico, anziché deterministico, la reazione $x_{\nu i}$ è una variabile casuale $X_{\nu i}$ caratterizzata da un distribuzione di probabilità $P(X_{\nu i})$ che, in generale, potrà dipendere dai parametri θ_ν e δ_i. Tali parametri in questo nuovo contesto - caratterizzato anche dall'errore e

dal caso come nella meccanica quantistica - diventano l'oggetto principale dell'inferenza, attraverso l'evidenza empirica costituita dalle reazioni osservate $x_{\nu i}$. I lavori di Rasch (1960, 1961 e 1977) favoriscono la scoperta della stretta connessione tra l'*oggettività specifica* da un lato e le *statistiche sufficienti* dall'altro, fino a giungere al risultato di Andersen (1977) il quale dimostra che i sistemi di riferimento deterministici caratterizzati dalla proprietà della *oggettività specifica* sono i soli che ammettono l'esistenza di statistiche sufficienti per i parametri, una volta trasposti in chiave probabilistica. Da questo consegue che sistemi di riferimento probabilistici caratterizzati da modelli $P(X_{\nu i})$ che ammettono statistiche sufficienti costituiscono la condizione necessaria e sufficiente per l'*oggettività specifica* del sistema di riferimento deterministico corrispondente. Per questo si dice anche che il modello di Rasch è un *Item Response Model* (cfr. Hambleton e Swaminathan, 1985) che cerca di misurare una o più variabili quantitative latenti sulla base di una scala di misura metrica, e che possiede le proprietà della *sufficienza*, *separabilità*, *oggettività specifica* e *additività latente*.

Il lavoro di Rasch, e successivamente quello di altri ricercatori (Wright, 1968, 1977; Andrich, 1978a, 1978b, 1978c; Linacre, 1989) porta a individuare una classe di modelli per la distribuzione di probabilità $P(X_{\nu i})$ che, ammettendo l'esistenza di statistiche sufficienti, sono in grado di assicurare la proprietà dell'*oggettività specifica*. Ma a questo punto viene in primo piano la questione relativa al fatto che gli elementi del sistema di riferimento deterministico $F = \{A, B, R\}$, sottostante il sistema di riferimento probabilistico, non devono contraddire il modello, affinché sia possibile che il processo inferenziale basato sulla osservazione delle reazioni $x_{\nu i}$ (realizzazione di una variabile casuale $X_{\nu i}$) porti a stime di θ_ν e δ_i che costituiscano "vere misure" (nel senso dell'*oggettività specifica*). In questo senso i modelli probabilistici di Rasch non sono, al contrario di quanto usualmente si può credere, solo strumenti statistici per la rappresentazione e la sintesi della realtà osservata, ma piuttosto una guida nella "scoperta" di sistemi di riferimento utili per misurare entità latenti nell'ambito di fenomeni di interesse che sono dipendenti dal contesto di osservazione e dal caso. Si noti che l'esistenza di statistiche sufficienti per questi modelli, oltre la proprietà della *oggettività specifica*, garantisce anche la

possibilità di ottenere stimatori con proprietà desiderabili come la correttezza e la consistenza, a patto di utilizzare metodi di stima adeguati (cfr. Hambleton e Swaminathan, 1985). Tali proprietà di correttezza e consistenza non sono invece garantite per altri modelli della classe *IRT* che nel "generalizzare" il RM perdono la proprietà della sufficienza portando la ricerca di misure su strade improduttive e prive di oggettività, oltre a presentare problemi di stima non facilmente risolvibili.

1.2.3 Gli assunti del Rasch Model (RM)

Tutti i modelli della famiglia di Rasch si fondano su tre assunti (Hambleton e Swaminathan, 1985):

A1. **Unidimensionalità**. Esiste una entità unidimensionale θ_ν, detta abilità latente, associata ad un generico soggetto ν, che determina la sua capacità di superare la prova a cui è sottoposto; le prove sono relative a tale dimensione unica e sono caratterizzate da una difficoltà δ_i con $i = 1, 2, \ldots, I$.

A2. **Monotonicità**. $P(X_{\nu i} > t|\theta_\nu, \delta_i)$ è una funzione monotòna della abilità θ_ν, per ogni i e ogni t. Soggetti con abilità più elevate hanno una maggiore probabilità di rispondere correttamente, di superare le prove o ricevere una valutazione elevata. Questo assunto consente di utilizzare il vettore delle osservazioni $\mathbf{X}_\nu = X_{\nu 1}, X_{\nu 2}, \ldots, X_{\nu I}$ relativo alle reazioni del soggetto ν alle diverse prove, come una serie di misure ripetute sullo stesso soggetto.

A3. **Indipendenza locale**. $P(\mathbf{X}_\nu|\theta_\nu, \delta_1, \delta_2, \ldots, \delta_I) = \prod_{i=1}^{I} P(X_{\nu i}|\theta_\nu, \delta_i)$, ossia, condizionatamente all'abilità del soggetto, le reazioni alle diverse prove sono indipendenti tra loro.

1.2.4 Il Modello di Rasch Dicotomico o Simple Logistic Model

Nel Modello di Rasch (RM più brevemente in seguito) Dicotomico si suppone che gli unici due parametri (entrambi rappresentabili sulla stessa scala di misura) che interagiscono per produrre il risultato aleatorio dicotomico $X_{\nu i}$, quando un soggetto ν risponde ad un item i, siano θ_ν, l'abilità latente del sog-

getto, e δ_i, la difficoltà incognita dell'item. Il modello matematico che governa la probabilità della variabile aleatoria $X_{\nu i}$ è il *Simple Logistic Model* (*SLM*), espresso dalla

$$Pr\{X_{\nu i} = 1; \theta_\nu, \delta_i\} = \frac{exp\{\theta_\nu - \delta_i\}}{1 + exp\{\theta_\nu - \delta_i\}} \tag{1.34}$$

nel caso di risposta corretta, e

$$Pr\{X_{\nu i} = 0; \theta_\nu, \delta_i\} = \frac{1}{1 + exp\{\theta_\nu - \delta_i\}} \tag{1.35}$$

nel caso di risposta non corretta.
Le due espressioni possono essere rappresentate nella forma sintetica che le comprende entrambe

$$Pr\{X_{\nu i}; \theta_\nu, \delta_i\} = \frac{exp\{X_{\nu i}(\theta_\nu - \delta_i)\}}{1 + exp\{\theta_\nu - \delta_i\}} \tag{1.36}$$

con $X_{\nu i}$ variabile casuale che può assumere solamente il valore 0 (quando la risposta è sbagliata) o il valore 1 (quando la risposta è corretta).

È immediato verificare che la probabilità di una risposta corretta ($X_{\nu i} = 1$) è uguale a 0.5 solo quando $\theta_\nu = \delta_i$, cioè quando l'abilità dell'individuo è uguale alla difficoltà dell'item. Tale proprietà è coerente con l'idea secondo la quale, se un soggetto incontra un item che presenta la stessa intensità relativa alla caratteristica da misurare, la probabilità che il soggetto "prevalga" sulla prova è uguale alla probabilità che quest'ultima "prevalga" sul soggetto.
Per un fissato δ_i, al variare di θ, si ottiene l'*Item Response Function* (*IRF*), conosciuta anche con il nome di *Item Characteristic Curve* (*ICC*), una curva logistica che cresce da 0 (per θ che tende a -∞), a 0.5 (quando $\theta = \delta_i$) fino a 1 (per θ che tende a +∞). Nel *RM* dicotomico tutte le *ICC* sono curve logistiche con la stessa pendenza e quindi parallele tra di loro; la sola caratteristica che le distingue è la posizione sulla scala di misura, determinata dalla difficoltà stimata dell'item.

Indicando con $\underline{\delta}$ il vettore a k dimensioni dei parametri degli item e con $\underline{\theta}$ il vettore a n dimensioni dei parametri degli individui, la probabilità dell'elemento $X_{\nu i}$ di ogni cella della matrice $n \times k$ $\mathbf{X}$, è espressa dalla formula (1.36).

Per ricavare la distribuzione di probabilità congiunta dell'intera matrice $\mathbf{X}$, si ipotizza:

1. che ci sia indipendenza nei vettori di risposta tra individui;

2. che, dato θ, le risposte di un individuo a item diversi siano stocasticamente indipendenti.

Secondo una possibile interpretazione di questa proprietà, conosciuta come indipendenza locale (*local independence*), non solo tutti gli item devono misurare la stessa variabile latente, ma ognuno deve, inoltre, fornire informazioni indipendenti riguardo la posizione dell'individuo relativamente a quella variabile.

Per ogni scelta di 0 e 1 nelle celle della matrice $\mathbf{X}$ si ottiene una probabilità data da

$$P(\mathbf{X} = \mathbf{x}; \underline{\theta}, \underline{\delta}) = \prod_{\nu=1}^{n} \prod_{i=1}^{k} \frac{exp[x_{\nu i}(\theta_\nu - \delta_i)]}{1 + exp(\theta_\nu - \delta_i)}. \tag{1.37}$$

Se tutti i parametri δ_i e θ_ν fossero noti, questa definirebbe una distribuzione di probabilità su tutte le matrici $n \times k$ con valori di ingresso $\{0, 1\}$ in ogni cella. In pratica però i parametri sono ignoti e devono essere stimati (con uno dei metodi che verranno spiegati in seguito); l'univocità delle stime è soddisfatta se si impone la restrizione di fissare un'origine arbitraria sulla scala.

Si è già detto che il RM è il solo modello che soddisfa l'*oggettività specifica* e che, di conseguenza, offre diversi vantaggi che gli altri metodi di stima di variabili latenti non hanno (come, ad esempio, l'Analisi Fattoriale Confirmatoria, che soddisfa l'unidimensionalità e produce misure su una scala ad intervallo). Uno dei vantaggi più apprezzabili è che esso ammette *statistiche sufficienti*: nella fattispecie il punteggio totale delle risposte corrette di ogni persona è una *statistica sufficiente* per la stima della sua abilità latente, e il punteggio totale di ogni item, ottenuto sommando le risposte corrette di tutti gli individui, è una statistica sufficiente per la stima della sua difficoltà ignota.

1.2.5 2PLM e 3PLM

Il *RM* fa parte, almeno dal punto di vista matematico, della più ampia famiglia dei Modelli dell'*Item Response Theory* che ipotizza che la risposta di un soggetto a un item dipenda sia dalla caratteristica del soggetto che dalle caratteristiche dell'item (Lord e Novick, 1968). Il *RM* è un modello logistico (la funzione che rappresenta la probabilità è una distribuzione logistica cumulata) a un parametro (*1PLM*) in quanto prevede un unico parametro per l' item: la sua difficoltà δ_i.

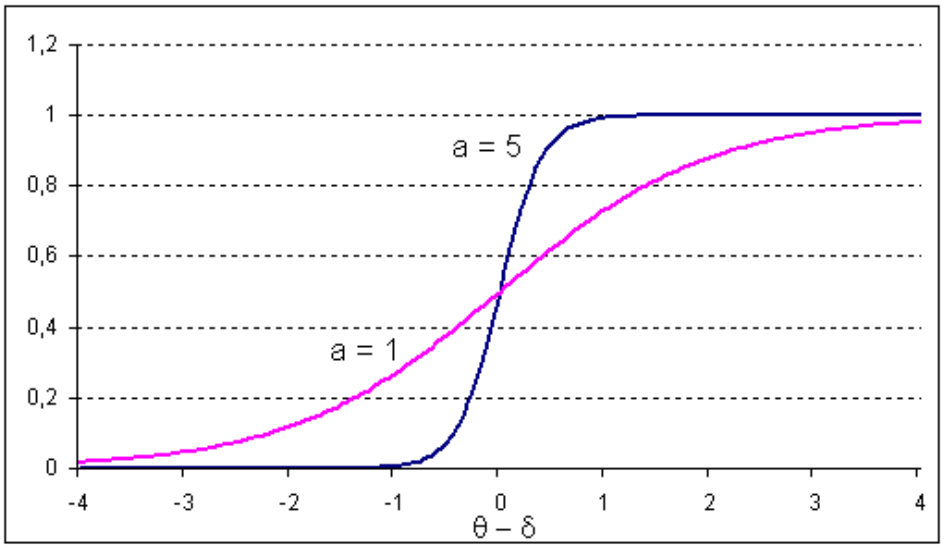


Figura 1.1: ICC con differenti item discrimination

Una seconda classe di modelli dell'*IRT* considera la possibilità che gli item differiscano tra di loro non solo per il livello di difficoltà ma anche per come discriminano tra i soggetti e introduce un secondo parametro a_i: l'*item discrimination*. Nel *2PLM* la probabilità di una risposta corretta è espressa dalla

$$Pr\{X_{\nu i} = 1; \theta_\nu, \delta_i, a_i\} = \frac{exp\{a_i(\theta_\nu - \delta_i)\}}{1 + exp\{a_i(\theta_\nu - \delta_i)\}} \tag{1.38}$$

La pendenza della curva in θ_ν è pari a:

$$a_i Pr\{X_{\nu i} = 1\}(1 - Pr\{X_{\nu i} = 1\})$$

che per $\theta_\nu = \delta_i$ vale 0.25 a_i. Quindi se a_i aumenta, aumenta anche la pendenza della curva. Nel caso limite in cui a_i cresca all'infinito, la *ICC* approssima una funzione a salti che vale 0 per $\theta_\nu < \delta_i$ e 1 per $\theta_\nu > \delta_i$ e che rappresenta la funzione di un item di Guttman con una discriminazione perfetta se $\theta_\nu = \delta_i$ e nessuna discriminazione per valori minori o maggiori di δ_i.

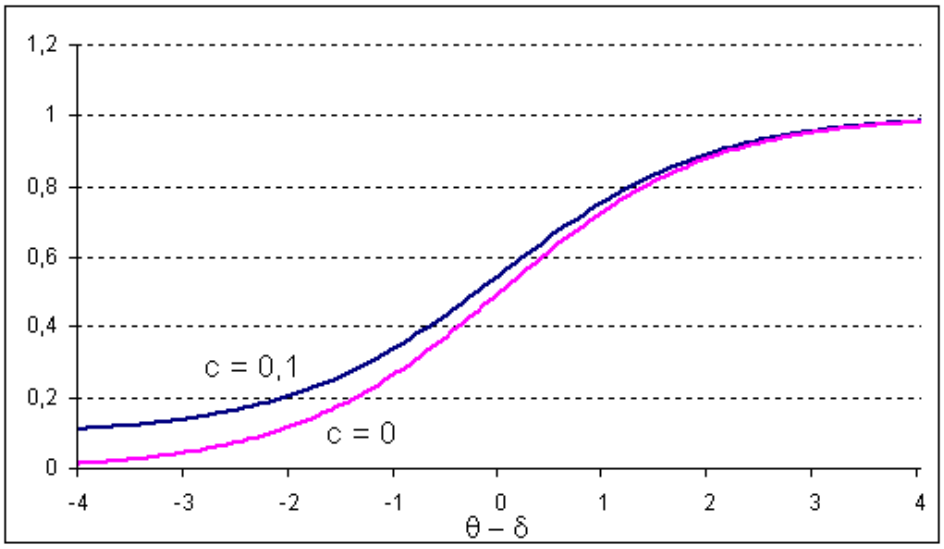


Figura 1.2: ICC con differenti parametri di guessing

Nel caso di domande a risposta multipla una terza classe di modelli dell'*IRT* prevede la possibilità che un soggetto, pur ignorando la risposta corretta, possa comunque rispondere esattamente affidandosi al caso (*guessing*). Con il *3PLM* la probabilità di rispondere correttamente diventa:

$$Pr\{X_{\nu i} = 1; \theta_\nu, \delta_i, a_i, c_i\} = c_i + (1 - c_i)\frac{exp\{a_i(\theta_\nu - \delta_i)\}}{1 + exp\{a_i(\theta_\nu - \delta_i)\}} \qquad (1.39)$$

dove c_i rappresenta l'asintoto per $\theta_\nu \to -\infty$, che, in generale, non coincide con il reciproco del numero di risposte dell'item ma è un parametro che va stimato assieme a δ_i e a_i. Il *2PLM* è un caso particolare del *3PLM* (quando $c_i = 0$) così come il *1PLM* è un caso particolare del *2PLM* (quando tutti i parametri di discriminazione sono supposti uguali). Il *RM*, a sua volta, è un caso particolare di *1PLM* in cui si ipotizza una discriminazione uguale per

tutti gli item e di valore unitario (la pendenza in $\theta_\nu = \delta_i$ vale sempre 0.25 per tutti gli item).

Credere che il *2PLM* e il *3PLM* siano strumenti di misura migliori rispetto al *1PLM* di Rasch per il fatto che prevedono, rispettivamente, uno o due ulteriori parametri che possono migliorare l'adattamento ai dati osservati è una falsa convinzione. Infatti, poiché questi due parametri extra non sono additivi, la proprietà fondamentale dell'*oggettività specifica* decade e nelle applicazioni pratiche si ottengono risultati paradossali: "Item discriminations increase without limit. Person abilities increase or decrease without limit" (Lord, 1968, pp.1015-1016), anche per dati generati per adattarsi esattamente al *3PLM*, "only item difficulty is satisfactorily recovered by [the 3P computer program] LOGIST." (Lord, 1975, p.13), "If restraints are not imposed, the estimated value of discrimination is likely to increase without limit" (Lord, 1975, p.14), "Left to itself, maximum likelihood estimation procedures would produce unacceptable values of guessing" (Lord, 1975, p.16), "During estimation in the two and three parameter models...the item parameter estimates drift out of bounds" (Swaminathan, 1983, p.34); "Range restrictions (must be) applied to all parameters except the item difficulties" to control "the problem of item discrimination going to infinity" (Wingersky, 1983, p.48); "Bias [in person measures] is significant when ability estimates are obtained from estimated item parameters...And, in spite of the fact that the calibration and cross-validation samples are the same for each setting, the bias differs by test." (Stocking, 1989, p.18), "Running LOGIST to complete convergence allows too much movement away from the good starting values" (Stocking, 1989, p.25).

Benjamin D. Wright (MESA Psychometric Laboratory, http://www.rasch.org/memo62.htm) spiega che: "The reason why 2P and 3P IRT models do not converge is clear in Birnbaum's original (Lord and Novick, 1968 pp.421-422) estimation equations":

$$\sum_i a_i x_{\theta i} = \sum_i a_i P_{\theta i} \rightarrow \theta$$

$$\sum_\theta \theta x_{\theta i} = \sum_\theta \theta P_{\theta i} \rightarrow a_i$$

"These equations are intended to iterate reciprocally to convergence. When

the first equation is applied to a person with a correct response $x_i = 1$ on an item with discrimination $a_i > 1$, their ability estimate is increased by the factor a_i. When the second equation is applied, the same person response $x_i = 1$ is multiplied by their increased ability estimate which further increases discrimination estimate a_i. The presence of response $x_i = 1$ on both sides of these reciprocal equations produces a feedback which soon escalates the estimates for item discrimination a_i and person measure to infinity."

Per tutte queste ragioni, sia di tipo computazionale, ma soprattutto di tipo teorico, legate all'idea della individuazione di un sistema di riferimento con la proprietà dell'*oggettività specifica*, si deve evidenziare la netta separazione, quasi di tipo dottrinale, tra gli utilizzatori del modello di Rasch e i fautori dei modelli *IRT* più in generale. C'è da sottolineare a favore della corrente di pensiero "Raschista", che gli studi che si ispirano a questa, spesso hanno lo scopo di validare un particolare sistema di riferimento per la misura di una variabile latente, nel tempo e nello spazio, ovvero quello di verificare la separabilità dei risultati del sistema di riferimento soprattutto per quanto concerne la prova: in molti lavori internazionali (vedi scala FIM, Tesio 1995) si cerca infatti di verificare se le difficoltà degli item siano le medesime in luoghi, tempi e rispetto a popolazioni differenti. Il che costituisce una forma di "controllo di qualità" delle scale stesse: caratteristica che gli altri modelli dell'*IRT* non possiedono.

Capitolo 2

I Modelli di Rasch

2.1 Metodi di stima dei parametri

2.1.1 Stima dei parametri degli item

Per analizzare i diversi metodi di stima dei parametri in situazioni generali più realistiche si suppone che nella matrice delle risposte $\mathbf{X}$ ci possano essere alcuni dati mancanti di sistema, come, ad esempio, nel caso in cui si vogliono agganciare assieme più test dove gruppi distinti di individui rispondono a gruppi distinti di item. Si introduce allo scopo la matrice $\boldsymbol{B}$, che si assume nota prima di somministrare la prova, in cui l'elemento $b_{\nu i}$ è posto uguale a 0 se l'item I_i non è somministrato all'individuo S_ν, e il corrispondente elemento $x_{\nu i}$ della matrice dei dati $\mathbf{X}$ uguale a un valore arbitrario a con $0 < a < 1$, mentre è posto uguale a 1 se l'item I_i è somministrato all'individuo S_ν. In generale, quindi, per le due matrici $\mathbf{X}$ e $\boldsymbol{B}$ si ha che:

$$x_{\nu i} = \begin{cases} 1 & \text{se } S_\nu \text{ risponde correttamente all'item } I_i \\ 0 & \text{se } S_\nu \text{ risponde in modo errato all'item } I_i \\ a & \text{se } I_i \text{ non è somministrato a } S_\nu \end{cases}$$

$$b_{\nu i} = \begin{cases} 1 & \text{se } I_i \text{ è somministrato a } S_\nu \\ 0 & \text{se } I_i \text{ non è somministrato a } S_\nu \end{cases}$$

In caso di matrice di dati completa $b_{\nu i} = 1$ per tutti i valori di ν e di i e le formule che seguono si semplificano. Per ogni cella della matrice dei dati $\mathbf{X}$ si ha, per $x_{\nu i} \in \{0,\, a,\, 1\}$ e il corrispondente $b_{\nu i} \in \{0, 1\}$:

$$Pr\{X_{\nu i} = x_{\nu i}; \theta_\nu, \delta_i\} = \frac{exp\{b_{\nu i} x_{\nu i}(\theta_\nu - \delta_i)\}}{\{1 + exp\{\theta_\nu - \delta_i\}\}^{b_{\nu i}}} \tag{2.1}$$

Nelle celle di $\mathbf{X}$ in cui il valore è mancante da sistema si ha, per definizione, $x_{\nu i} = a$ mentre per gli altri casi le formule si riconducono alla 1.36. Per l'ipotesi di indipendenza locale e di indipendenza tra i soggetti, segue che, per ogni scelta di 0 e 1 nelle celle osservate della matrice $\mathbf{X}$, la probabilità congiunta è

$$Pr\left\{\mathbf{X} = \mathbf{x}; \underline{\theta}, \underline{\delta}\right\} = \prod_{\nu=1}^{n} \prod_{i=1}^{k} \frac{exp\left\{b_{\nu i} x_{\nu i}(\theta_\nu - \delta_i)\right\}}{\left\{1 + exp\left\{\theta_\nu - \delta_i\right\}\right\}^{b_{\nu i}}} \tag{2.2}$$

Ancora una volta, se tutti i parametri fossero noti, questa definirebbe una distribuzione di probabilità su tutte le n x k matrici con valori appartenenti all'insieme $\{0, 1\}$ per le risposte osservate, e pari ad a per tutte le osservazioni mancanti. In questo contesto tutti i parametri δ_i degli item e tutti i parametri θ_ν degli individui sono ignoti e devono essere stimati; nei paragrafi che seguono ci si soffermerà sui principali metodi di stima dei k parametri degli item mentre gli n parametri degli individui agiscono come parametri "di disturbo".

In letteratura esistono diversi metodi per la stima dei parametri dei modelli dell'*IRT* (Hambleton R.K. e Swaminathan H., 1985; Hambleton R. K., Swaminathan H. e Rogers H. J., 1991; Crocker L. M. e Algina J., 1986). Per quanto riguarda i metodi di stima delle difficoltà degli item si possono considerare due criteri di classificazione. Il primo criterio concerne i parametri degli individui $\underline{\theta}$ che possono:

- essere stimati congiuntamente con i parametri degli item;
- essere eliminati condizionando le probabilità di risposta rispetto a una loro statistica sufficiente;
- essere marginalizzati mediante una procedura di integrazione.

Il secondo criterio riguarda il metodo di stima vero e proprio che si utilizza per stimare i parametri degli item. Generalmente si utilizzano metodi di massima verosimiglianza (congiunta, condizionata, marginale) ma esistono anche altre procedure alternative come le *Stime Congiunte* e *Marginali* di tipo *Bayesiano* (*Joint and Marginal Bayesian Estimation*), *Stime Euristiche* (*Heuristic Estimation*) e *Stime basate sull'Analisi Fattoriale*.
La *Stima di Massima Verosimiglianza Congiunta* (*Joint Maximum Likelyhood Estimation*, *JMLE*), la *Stima di Massima Verosimiglianza Condizionata* (*Conditional Maximum Likelyhood Estimation*), la *Stima di Massima Verosimiglianza Marginale* (*Marginal Maximum Likelyhood Estimation*, *MMLE*) e la *Stima per Dati Appaiati* (*Pair-Wise Parameter Estimation*, *PWPE*) sono i metodi di stima più diffusi. La *JMLE* e la *MMLE* possono essere impiegate nella stima dei parametri del *1PLM*, *2PLM* e *3PLM*; la *CMLE* e la *PWPE*, invece, si applicano esclusivamente al *RM*.

2.1.2 La Massima Verosimiglianza Congiunta (*JML*)

Dall'equazione (2.2) si ricava che la log-verosimiglianza non condizionata o congiunta è espressa dalla

$$\ln L(\underline{\theta},\underline{\delta}) = \sum_{\nu=1}^{n}\sum_{i=1}^{k} b_{\nu i}x_{\nu i}(\theta_\nu - \delta_i) - \sum_{\nu=1}^{n}\sum_{i=1}^{k} b_{\nu i}\ln(1+exp(\theta_\nu - \delta_i))$$

$$= \sum_{\nu=1}^{n} x_{\nu.}\theta_\nu - \sum_{i=1}^{k} x_{.i}\delta_i - C(\underline{\theta},\underline{\delta}) \tag{2.3}$$

dove $C(\underline{\theta},\underline{\delta})$ non dipende dai dati osservati e le quantità

$$x_{.i} = \sum_{\nu=1}^{n} b_{\nu i}x_{\nu i} \quad \text{e} \quad r_\nu = x_{\nu.} = \sum_{i=1}^{k} b_{\nu i}x_{\nu i}$$

sono, rispettivamente, i totali di colonna e di riga dei valori osservati della matrice $\mathbf{X}$ che costituiscono le statistiche sufficienti per stimare, rispettivamente, i k parametri degli item e gli n parametri dei soggetti.
Calcolando le derivate della (2.3) si ottiene che le stime di *JML* soddisfano i

due seguenti gruppi di equazioni:

$$x_{.i} = \sum_{\nu=1}^{n} \frac{b_{\nu i}\, exp(\theta_\nu - \delta_i)}{1 + exp(\theta_\nu - \delta_i)} \quad \text{i} = 1, \ldots , \text{k} \tag{2.4}$$

$$r_\nu = \sum_{i=1}^{k} \frac{b_{\nu i}\, exp(\theta_\nu - \delta_i)}{1 + exp(\theta_\nu - \delta_i)} \quad \nu = 1, \ldots , \text{n} \tag{2.5}$$

Le 2.4 e 2.5 esprimono l'uguaglianza tra i valori delle statistiche sufficienti e i loro valori attesi.

Il sistema di $n + k$ equazioni si riduce quando due o più gruppi di persone o due o più gruppi di item presentano lo stesso *score*. Gli *score* delle persone possono assumere valori pari a *0, 1 , . . . , k* ma gli *zero scores* (risposte tutte sbagliate) e i *perfect scores* (risposte tutte corrette) ricoprono un ruolo particolare in quanto per questi valori, con questo metodo di stima, non è possibile ottenere stime finite, mentre delle k equazioni relative agli item espresse dalla 2.4 una è superflua per la restrizione dovuta alla normalizzazione. Così, per dati completi, ci sono al più $2k$ - 2 equazioni indipendenti, e un numero minore nel caso in cui alcuni *score* degli individui non sono stati osservati e/o quando alcuni item presentano *score* identici. Inoltre, se un item è sbagliato da tutte le persone a cui è somministrato ($x_{.i} = 0$), la soluzione formale della 2.4 diverge a $\delta_i = + \infty$; così se un item è superato da tutte le persone a cui è somministrato ($x_{.i} = n$), la soluzione formale diverge a $\delta_i = - \infty$. In questi due casi si può pensare che i dati non forniscono informazioni sufficienti per localizzare le difficoltà degli item sulla scala.
Un discorso analogo vale per i parametri degli individui; se un individuo non dà nessuna risposta corretta, $r_\nu = 0$ e la soluzione formale dell'equazione contenuta nella 2.5 è $\theta_\nu = - \infty$, mentre se egli risponde in modo esatto a tutti gli item, $r_\nu = \text{k}$ e la soluzione formale è $\theta_\nu = + \infty$. Nell'applicazione pratica del metodo item e individui con *zero score* o *perfect score* vengono estromessi dalla procedura di stima della *JML*.
Anche quando si assume che la probabilità dei punteggi nulli o perfetti è molto piccola, una soluzione finita delle equazioni di stima non è sempre possibile (Fischer, 1974, 261-263, e Fischer, 1981, 71-72). Per esempio nel caso in cui

gli item possono essere divisi in due gruppi in modo tale che tutti gli individui danno risposte positive (o missing) a tutti gli item del primo gruppo, e risposte negative (o missing) a tutti gli item del secondo gruppo, gli item del secondo gruppo sembrano essere molto più difficili degli item del primo gruppo, e non è possibile effettuare nessun confronto dei parametri degli item dei due gruppi. Nella letteratura dati siffatti sono chiamati *ill-conditioned*, contrapposti a quelli denominati *well-conditioned* che non presentano sottogruppi di tale genere.
La stima di *JLM* procede con un semplice processo di iterazione. I valori iniziali sono aggiornati o alternando la soluzione della 2.4 e della 2.5, o aggiornando simultaneamente tutti i parametri degli item e delle persone mediante una routine che massimizza la 2.3. Ad ogni modo nella *JML* la stima dei parametri degli item dipende dalla stima dei parametri delle persone e viceversa.

Il maggior inconveniente del metodo di Massima Verosimiglianza Congiunta è che fornisce stime che non sono consistenti per $n \to \infty$, con k fissato, sebbene la consistenza sussista nel caso in cui $n \to \infty$, $k \to \infty$, $n/k \to \infty$ (Andersen, 1973b, Haberman, 1977). Nella maggior parte dei casi pratici un campione molto ampio di persone viene valutato mediante un set limitato di item, nel cui caso il numero n di parametri di disturbo θ_ν eccede di gran lunga il numero k di parametri strutturali δ_i. Quindi le stime di *JML* degli item non sono mai consistenti, e neppure asintoticamente corrette; studi di simulazione hanno dimostrato che, per dati completi, il fattore correttivo $(k-1)/k$ rimuove gran parte del *bias* che comunque diventa trascurabile per test composti da molte domande.

2.1.3 La Massima Verosimiglianza Condizionata (*CML*)

Per una proprietà delle famiglie esponenziali la distribuzione condizionata, data una statistica sufficiente per i parametri di disturbo, non dipende più da questi ultimi. Nel *RM* la massimizzazione della log-verosimiglianza condizionata, dati i *total score* degli individui $r_\nu = x_{\nu.}$, conduce a stime dei parametri degli item migliori di quelle ricavate con il metodo della *JML*.

Ponendo $\xi_\nu = exp(\theta_\nu)$ e $\epsilon_i = exp(-\delta_i)$, la (1.34) e la (1.35) si riscrivono:

$$Pr\{X_{\nu i} = 1; \theta_\nu, \delta_i\} = \frac{\xi_\nu \epsilon_i}{1 + \xi_\nu \epsilon_i} \tag{2.6}$$

e

$$Pr\{X_{\nu i} = 0; \theta_\nu, \delta_i\} = \frac{1}{1 + \xi_\nu \epsilon_i}. \tag{2.7}$$

Nel caso più semplice in cui ci sono due soli item ($i = 2$) e i dati sono completi, fissato un valore per l'abilità ξ_ν, si ottiene:

$$Pr\{X_{\nu 1} = 1, X_{\nu 2} = 0 | X_{\nu \cdot} = r_\nu = 1\} = \frac{\frac{\xi_\nu \epsilon_1}{(1 + \xi_\nu \epsilon_1)(1 + \xi_\nu \epsilon_2)}}{\frac{\xi_\nu \epsilon_1 + \xi_\nu \epsilon_2}{(1 + \xi_\nu \epsilon_1)(1 + \xi_\nu \epsilon_2)}} = \frac{\epsilon_1}{\epsilon_1 + \epsilon_2} \tag{2.8}$$

e simmetricamente

$$Pr\{X_{\nu 1} = 0, X_{\nu 2} = 1 | X_{\nu \cdot} = r_\nu = 1\} = \frac{\epsilon_2}{\epsilon_1 + \epsilon_2} \tag{2.9}$$

. Nella 2.8 (2.9) la probabilità che un individuo a cui sono sottoposti due item risponda correttamente al primo (secondo) e in modo sbagliato al secondo (primo), subordinatamente al fatto che egli risponda correttamente ad un item, non dipende dalla sua abilità. Così il rapporto tra la 2.8 e la 2.9, uguale a $\frac{\epsilon_1}{\epsilon_2}$, e, in generale, la verosimiglianza condizionata dipendono solamente dai parametri degli item.
Nel caso di tre item ($i = 3$) e dati completi si ottiene

$$\begin{aligned}
Pr\{X_{\nu 1} = 1, X_{\nu 2} = 0, X_{\nu 3} = 0, | X_{\nu \cdot} = r_\nu = 1\} &= \frac{\epsilon_1}{\epsilon_1 + \epsilon_2 + \epsilon_3} \\
Pr\{X_{\nu 1} = 0, X_{\nu 2} = 1, X_{\nu 3} = 0, | X_{\nu \cdot} = r_\nu = 1\} &= \frac{\epsilon_2}{\epsilon_1 + \epsilon_2 + \epsilon_3} \\
Pr\{X_{\nu 1} = 0, X_{\nu 2} = 0, X_{\nu 3} = 1, | X_{\nu \cdot} = r_\nu = 1\} &= \frac{\epsilon_3}{\epsilon_1 + \epsilon_2 + \epsilon_3}
\end{aligned}$$

e

$$Pr\{X_{\nu 1} = 1, X_{\nu 2} = 1, X_{\nu 3} = 0, | X_{\nu \cdot} = r_\nu = 2\} = \frac{\epsilon_1 \epsilon_2}{\epsilon_1 \epsilon_2 + \epsilon_2 \epsilon_3 + \epsilon_1 \epsilon_3}$$

$$Pr\{X_{\nu 1}=1, X_{\nu 2}=0, X_{\nu 3}=1, |X_{\nu\cdot}=r_\nu=2\} = \frac{\epsilon_1\epsilon_3}{\epsilon_1\epsilon_2+\epsilon_2\epsilon_3+\epsilon_1\epsilon_3}$$

$$Pr\{X_{\nu 1}=0, X_{\nu 2}=1, X_{\nu 3}=1, |X_{\nu\cdot}=r_\nu=2\} = \frac{\epsilon_2\epsilon_3}{\epsilon_1\epsilon_2+\epsilon_2\epsilon_3+\epsilon_1\epsilon_3}$$

Da queste si ricava che la probabilità che l'item 1 sia corretto, subordinatamente ad un *individual score* uguale a 2 è:

$$Pr\{X_{\nu 1}=1, X_{\nu 2}=1, X_{\nu 3}=0, |r_\nu=2\}+$$
$$+ Pr\{X_{\nu 1}=1, X_{\nu 2}=0, X_{\nu 3}=1, |r_\nu=2\} = \frac{\epsilon_1\epsilon_2+\epsilon_1\epsilon_3}{\epsilon_1\epsilon_2+\epsilon_2\epsilon_3+\epsilon_1\epsilon_3}$$

Nel caso generale di k item e n persone, con la possibilità che i dati non siano completi, la verosimiglianza condizionata per la matrice dei dati $\mathbf{X}$ è (Fischer e Molenaar, 1995):

$$L(\epsilon|r) = \prod_{\nu=1}^{n}\left(\prod_{i=1}^{k}\epsilon_i^{x_{\nu i}b_{\nu i}}\gamma_{r_\nu}^{-1}\right) = \left(\prod_{\nu=1}^{n}\gamma_{r_\nu}\right)^{-1}\prod_{i=1}^{k}\epsilon_i^{x_{\cdot i}} \tag{2.10}$$

dove la generica funzione elementare simmetrica γ_{r_ν} delle variabili $\epsilon_i b_{\nu i}$ è definita come la somma di tutti i prodotti dei r_ν:

$$\gamma_{r_\nu}(\epsilon_1 b_{\nu 1}, \ldots, \epsilon_k b_{\nu k}) = \sum_{y|r_\nu}\prod_{i=1}^{k}(\epsilon_i b_{\nu i})^{y_i}$$

con la sommatoria che varia tra tutti i pattern di risposta $y=(y_1,\ldots,y_k)$ con

$$\sum_{i=1}^{k} y_i b_{\nu i} = r_{\nu\cdot}$$

Nel caso di dati completi si ha che:

$$\gamma_0 = 1,$$

$$\gamma_1 = \epsilon_1+\epsilon_2+\ldots+\epsilon_k,$$

$$\gamma_2 = \epsilon_1\epsilon_2+\epsilon_1\epsilon_3+\ldots+\epsilon_{k-1}\epsilon_k,$$

$$\vdots$$

$$\gamma_k = \epsilon_1 \epsilon_2 \ldots \epsilon_k.$$

Nel caso di matrice di dati incompleta le ϵ_i vengono sostituite da $b_{\nu i}\epsilon_i$.

Le stime di *CML* si trovano massimizzando la 2.10; quindi ponendo uguale a zero tutte le sue derivate rispetto a ϵ_i, per $i = 1, \ldots, k$, si ricava la

$$\frac{\partial \gamma_{r_\nu}(b_{\nu 1}\epsilon_1, \ldots, b_{\nu k}\epsilon_k)}{\partial \epsilon_i} = b_{\nu i}\gamma^{(i)}_{r_\nu - 1} \tag{2.11}$$

dove $\gamma^{(i)}_{r_\nu - 1}$ indica la funzione simmetrica elementare di $r_\nu - 1$ argomenti $\epsilon_j b_{\nu j}$ senza $\epsilon_i b_{\nu i}$. Dopo alcuni passaggi algebrici si ottengono le equazioni che devono essere soddisfatte nella stima della *CML*:

$$x_{.i} - \sum_{\nu=1}^{n} \frac{\epsilon_i b_{\nu i}\gamma^{(i)}_{r_\nu - 1}}{\gamma_{r_\nu}} = 0 \quad \text{per } i = 1, \ldots, \text{k.} \tag{2.12}$$

In questo caso le statistiche sufficienti dei parametri degli item ($x_{.i}$) eguagliano la somma delle probabilità condizionate di rispondere correttamente agli item dato un punteggio totale di risposte corrette pari a r_ν. Il metodo della *CML* massimizza la verosimiglianza condizionata, dati i *total score* degli individui, e non la verosimiglianza totale. In generale tale condizione implica una perdita di informazione; infatti ci si può aspettare che anche la distribuzione dei *total score* contenga delle informazioni sui parametri degli item. Una diretta conseguenza di tale perdita di informazione è che, sebbene la matrice di varianza-covarianza asintotica stimata si possa determinare dalla matrice delle derivate seconde sostituendo i parametri con le loro stime, non vale più il limite di Cramer-Rao per la varianza asintotica degli stimatori (Fischer, 1974). Le stime di *CML* sono consistenti per $n \to \infty$, con k fissato, ma la procedura di stima è spesso lunga e complessa dal punto di vista computazionale (Andersen, 1973c).

2.1.4 La Massima Verosimiglianza Marginale (*MML*)

Con la Massima Verosimiglianza Marginale i parametri di disturbo θ_ν vengono eliminati mediante un processo di integrazione. La probabilità assoluta

di osservare una determinazione di $\mathbf{X}$ è espressa dalla

$$P(\mathbf{X} = \mathbf{x}|\underline{\theta}, \underline{\delta}) = P(\mathbf{X} = \mathbf{x}|r, \underline{\delta})P(R = r|\underline{\theta}, \underline{\delta}). \tag{2.13}$$

Nella *CML* si massimizza il primo fattore a destra della 2.13, trascurando il secondo fattore. Nella *MML*, invece, si stima anche il secondo fattore assumendo l'esistenza di una distribuzione per $\underline{\theta}$.

Sia $G(\theta)$ la funzione di ripartizione dell'abilità degli individui nella popolazione e si supponga che i parametri degli individui osservati siano un campione casuale di tale distribuzione. Sia S_ν un individuo di abilità θ_ν e sia x_ν qualsiasi vettore k-dimensionale con ingressi uguali ad, a con $0 < a < 1$, per gli item per cui $b_{\nu i} = 0$, e valori presi in modo arbitrario nell'insieme $\{0, 1\}$, per gli altri ingressi. Allora la probabilità di osservare il pattern di risposte x_ν è:

$$P(X_\nu = x_\nu|G, \delta) = \int_{-\infty}^{+\infty} \prod_{i=1}^{k} \frac{exp[b_{\nu i}x_{\nu i}(\theta - \delta_i)]}{[1 + exp(\theta - \delta_i)]^{b_{\nu i}}} dG(\theta). \tag{2.14}$$

La Verosimiglianza Marginale (funzione di G e di $\underline{\delta}$) di ottenere la matrice $\mathbf{X}$ dei dati osservati è data dal prodotto, per $\nu = 1, \ldots, n$, di questi integrali.
Naturalmente G non è nota. Se i dati osservati sono sufficientemente numerosi è possibile ricavare informazioni utili circa le proprietà della distribuzione di θ. Se, al contrario, i dati osservati sono scarsi si ipotizza che G appartenga a una determinata famiglia parametrica con pochi iperparametri incogniti (che sono stimati congiuntamente a $\underline{\delta}$); di solito si assume che θ abbia una distribuzione normale con media e varianza ignote sebbene tale scelta renda la procedura di stima alquanto laboriosa. Per superare quest'ultimo problema si può approssimare la distribuzione normale con una distribuzione discreta, con q classi di abilità latenti. Allora, indicate con

$$g(\theta_l) \quad l = 1, \ldots, q$$

le frequenze relative di θ_l, la probabilità di un pattern di risposte $\mathbf{x}$ diventa:

$$P(\mathbf{x}|g,\delta) = \sum_{l=1}^{q} P(\mathbf{x}|\theta_l)g(\theta_l). \tag{2.15}$$

Un'altra semplificazione si ottiene ipotizzando una distribuzione Normale Standardizzata per le abilità latenti; in questo caso l'origine della scala è fissata dalla distribuzione dei parametri degli individui.

Il metodo di *MML* offre vantaggi che quello di *CML* non ha. La *MML* fornisce stime di parametri finite anche per individui con *zero score* o *perfect score*, cosicché tali individui non devono essere rimossi dal processo di stima; infatti questi valori sono utili per trovare la distribuzione delle abilità, sebbene essi non forniscano alcuna informazione sulla posizione relativa dei parametri degli item. Se l'obiettivo è trovare la distribuzione delle abilità degli individui, la procedura della MML è manifestamente superiore.
Il rovescio della medaglia è dato dal fatto che la bontà delle stime dei parametri degli item è sensibilmente influenzata dalla scelta di *G*. La *MML* richiede di stimare o di ipotizzare una distribuzione della variabile latente; se tale ipotesi è sbagliata, le stime di *MML* possono risultare ben peggiori di quelle degli altri due metodi (come rileva Glas - 1989 - anche assumendo una distribuzione normale, le stime di *MML* possono essere molto distorte se la distribuzione dei parametri delle persone della popolazione da cui viene estratto il campione si scosta molto dalla normalità).
De Leeuw e Verhelst (1986) hanno dimostrato che la *CML* e la *MML* non parametrica sono asintoticamente (per n $\rightarrow +\infty$ e k fissato) equivalenti.

2.1.5 La Stima per Dati Appaiati (*PWPE*)

A partire dai primi anni 70 (Fischer e Scheiblechner, 1970), con lo scopo di evitare i lunghi tempi di calcolo richiesti dai metodi *JML* e *MML*, sono state sviluppate procedure alternative per la stima dei parametri degli item. Le più diffuse si basano sul confronto delle risposte osservate in ciascuna coppia di item, indipendentemente dalle risposte date a tutti gli altri item. Fischer (Fischer, 1974) ha proposto un metodo di ottimizzazione basato sul chi-quadrato

di Pearson (*Minchi estimation*). Successivamente Choppin (1983) e Van der Linden e Eggen (1986) hanno proposto un metodo simile, ma basato sulla pseudo-verosimiglianza, che sfrutta la proprietà dell'indipendenza condizionale; la *Stima per Dati Appaiati* (Andrich, 1988b), implementata in RUMM2020, è uno sviluppo di quest'ultimo metodo.

Si considerino due item diveri I_i e I_j. La

$$Pr\left\{(X_{\nu i}, X_{\nu j})|r = 1\right\} = \frac{exp(-X_{\nu i}\delta_i - X_{\nu j}\delta_j)}{[exp(-\delta_i) + exp(-\delta_j)]} \tag{2.16}$$

generalizza la probabilità che $X_{\nu i} = 1$ o che $X_{\nu j} = 1$, subordinatamente all'evento che la risposta ad un item sia corretta e la risposta all'altro item sia sbagliata. Se N individui hanno un punteggio $r = 1$, per l'indipendenza stocastica tra gli individui, segue che:

$$\begin{aligned}\lambda &= \prod_{\nu} Pr\left\{(X_{\nu i}, X_{\nu j})|r = 1\right\} = \\ &= \prod_{\nu} \frac{exp(-X_{\nu i}\delta_i - X_{\nu j}\delta_j)}{[exp(-\delta_i) + exp(-\delta_j)]^N} \\ &= \frac{exp[-\sum_{\nu} X_{\nu i}\delta_i - \sum_{\nu} X_{\nu j}\delta_j]}{[exp(-\delta_i) + exp(-\delta_j)]^N}.\end{aligned}$$

Dopo aver posto $s_i = \sum_{\nu} X_{\nu i}$ e $s_j = \sum_{\nu} X_{\nu j}$ (il numero totale di volte in cui è data una risposta corretta all'item i e j, rispettivamente, l'espressione diventa:

$$\lambda = \frac{exp[-s_i\delta_i - s_j\delta_j]}{[exp(-\delta_i) + exp(-\delta_j)]^N}. \tag{2.17}$$

Passando al logaritmo si ottiene

$$ln\lambda = -s_i\delta_i - s_j\delta_j - Nln[exp(-\delta_i) + exp(-\delta_j)]. \tag{2.18}$$

Ora, indicando con F_{ij} il numero di persone che hanno risposto positivamente a un item e in modo errato all'altro item, con f_{ij} il numero di persone che hanno risposto positivamente solo all'item I_i e con f_{ji} il numero di persone che hanno risposto positivamente solo all'item I_j, si ha che $F_{ij} = f_{ij} + f_{ji}$. Risulta

che per ogni coppia (i, j), la probabilità di osservare f_{ij} su F_{ij} è:

$$Pr\left\{(f_{ij}, f_{ji}); (\delta_i, \delta_j) | r = 1\right\} = \frac{F_{ij}! exp(-\delta_i)^{f_{ij}} exp(-\delta_j)^{f_{ji}}}{f_{ij}! f_{ji}! [exp(-\delta_i) + exp(-\delta_j)]^{F_{ij}}} \tag{2.19}$$

mentre la probabilità dell'intera matrice per dati appaiati risulta essere

$$\Lambda = (\prod_i \prod_{i \neq j} \prod_j \frac{F_{ij}!}{f_{ij}! f_{ji}!}) \frac{exp(-\sum_i \sum_j \delta_i f_{ij}) exp(-\sum_i \sum_j \delta_j f_{ji})}{\prod_i \prod_j [exp(-\delta_i) + exp(-\delta_j)]^{F_{ij}}} \tag{2.20}$$

Passando al logaritmo si ottiene:

$$ln\Lambda = C - \sum_i \sum_j f_{ij}\delta_i - \sum_i \sum_j f_{ji}\delta_j - \sum_i \sum_j F_{ij} \left\{\ln[exp(-\delta_i) + exp(-\delta_j)]\right\} \quad \text{con } i \neq j. \tag{2.21}$$

Dopo aver calcolato le derivate di $ln\Lambda$ e averle poste uguali a zero, si ottengono le equazioni delle soluzioni:

$$-s_i + \sum_j F_{ij}\hat{\pi}_{ij} = 0 \quad i = 1, \ldots, I \tag{2.22}$$

dove $s_i = \sum_j f_{ij}$ è il numero totale di volte in cui si risponde correttamente all'item e negativamente a tutti gli altri item. Anche in questo caso bisogna imporre la condizione $\sum_i \hat{\delta}_i = 0$ affinchè le stime siano univocamente determinate e definire un'origine della scala di misura.

La procedura di *Stima per Dati Appaiati* si basa su un'approssimazione della Massima Verosimiglianza poiché le f_{ij} che compaiono nella 2.19 e seguenti vengono trattate come se fossero indipendenti mentre in realtà non lo sono in quanto in coppia diverse di item ci possono essere gli stessi soggetti. Ciononostante gli stimatori che si ricavano, quando esistono, sono molto soddisfacenti: sono consistenti e simili, in efficienza, a quelli ricavati dai metodi *CML* e *MML* (Zwinderman, 1995).

2.1.6 Altri metodi di stima

Esistono altri metodi di stima che utilizzano le frequenze f_{ij}, delle persone che hanno risposto correttamente all'item I_i e in maniera errata all'item I_j, ma che non si basano sulla funzione di verosimiglianza (Fischer, 1974). In tutti questi metodi, che si basano sul fatto che il rapporto f_{ij}/f_{ji} è una stima di ϵ_i/ϵ_j, non è possibile determinare l'errore standard asintotico, ma le stime sono facili da calcolare e forniscono dei soddisfacenti valori iniziali da utilizzare nei metodi *JML*, *CML* e *MML*.

Un procedimento sfrutta il vincolo $\sum_j \delta_j = 0$ che implica $\prod_j \epsilon_j = 1$. Moltiplicando questi rapporti per tutti i $j \neq i$ si ottiene:

$$\prod_{j \neq i} \frac{f_{ij}}{f_{ji}} \cong \epsilon_i^k. \tag{2.23}$$

Quando tutte le f_{ij} e f_{ji} sono diverse da zero (cosa che succede raramente se il campione è di numerosità ridotta) la stima delle ϵ_i si ricava determinando la radice k-esima del termine di sinistra della 2.23.

Una seconda procedura minimizza la

$$\sum_i \sum_{j<i} (\frac{f_{ij}}{\epsilon_i} - \frac{f_{ji}}{\epsilon_j})^2$$

riconducendo la sua soluzione ad un problema di *eigenvalue* di una matrice k x k.

Il terzo metodo utilizza la procedura *Minchi* (termine preferito al "minimo chi-quadrato" poiché ogni addendo della 2.24 si distribuisce asintoticamente come una v.c. chi-quadrato con 1 gdl, ma gli addendi non sono indipendenti tra di loro per la sovrapposizione degli individui nelle diverse coppie di item) che minimizza la

$$\sum_i \sum_{j<i} \frac{(f_{ij}\epsilon_j - f_{ji}\epsilon_i)^2}{(f_{ij} + f_{ji})\epsilon_i\epsilon_j} \tag{2.24}$$

omettendo dal calcolo tutte le coppie (i, j) per cui vale $f_{ij} + f_{ji} = 0$. Le equazioni che si ricavano ponendo tutte le derivate parziali della 2.24 uguali a

zero sono (Fischer, 1974)

$$\epsilon_h^{-2} = \frac{\sum_i y_{ih}^2 \epsilon_i^{-1}}{\sum_j y_{hj}^2 \epsilon_j}$$

in cui

$$y_{ij} = \frac{f_{ij}}{\sqrt{f_{ij} + f_{ji}}}.$$

La convergenza si raggiunge rapidamente e le stime che si trovano sono molto simili a quelle della *CML*.
Poiché il rapporto f_{ij}/f_{ji} è uno stimatore consistente di ϵ_i/ϵ_j tutti e tre i metodi sono consistenti. Tutti e tre i metodi inoltre si applicano anche in presenza di dati mancanti.

2.2 Generalizzazioni dell'SLM

La caratteristica principale del *RM*, strettamente legata all'*oggettività specifica*, è l'*invarianza dei confronti* (*invariance of comparison*), in base alla quale il confronto fra item risulta indipendente dagli individui a cui sono somministrati e, in modo simmetrico, il confronto fra individui risulta indipendente dagli item utilizzati. Georg Rasch definì questa proprietà (Rasch, 1960) nel *Teorema di Separabilità*, formalizzato successivamente da Andersen e Olsen (Andersen e Olsen, 2001) in questi termini:

È possibile stimare i parametri degli item, $\delta_1, \ldots, \delta_k$, in una distribuzione in cui i parametri dei soggetti, $\theta_1, \ldots, \theta_n$, sono stati eliminati. Simmetricamente, è possibile stimare i parametri dei soggetti in una distribuzione in cui i parametri degli item sono stati eliminati. Il controllo del modello, infine, può essere basato su una distribuzione in cui entrambi i parametri sono stati eliminati.

Nel 1961, durante il "Quarto Simposio di Berkley sulla Teoria della Probabilità e la Statistica Matematica", Rasch propose due classi di modelli che godevano di tale proprietà. La prima classe si presenta nella seguente formula:

$$P\{X_{\nu i};\theta_\nu,\delta_i\} = \frac{exp(\phi_x\theta_\nu + \psi_x(-\delta_i) + \chi_x\theta_\nu\delta_i + \kappa_x)}{\sum_{k=0}^{m} exp(\phi_k\theta_\nu + \psi_k(-\delta_i) + \chi_k\theta_\nu\delta_i + \kappa_k)} \tag{2.25}$$

dove ϕ e ψ rappresentano funzioni punteggio delle categorie, χ e κ vettori di costanti e $X_{\nu i} = x$ la risposta dell'individuo ν in una qualunque delle $m+1$ categorie dell'item i. All'interno di tale classe, solo nel caso in cui $\chi \equiv 0$, è possibile individuare statistiche sufficienti per la stima dei parametri degli individui indipendentemente dagli item e viceversa. Con questa imposizione il modello si riduce a

$$P\{X_{\nu i};\delta_i,\theta_\nu\} = \frac{exp(\phi_x\theta_\nu + \psi_x(-\delta_i) + \kappa_x)}{\sum_{k=0}^{m} exp(\phi_k\theta_\nu + \psi_k(-\delta_i) + \kappa_k)}. \tag{2.26}$$

Il *SLM* ne è un caso particolare.

Rasch ricavò la seconda classe di modelli eliminando dalla 2.26 uno dei due parametri scalari, ottenendo così

$$P\{X_{\nu i};\delta_i,\theta_\nu\} = \frac{exp(\phi_x(\theta_\nu - \delta_i) + \kappa_x)}{\sum_{k=0}^{m} exp(\phi_k(\theta_\nu - \delta_i) + \kappa_k)} \tag{2.27}$$

La 2.27 è alla base delle estensioni del *SLM*. Andersen (1977) dimostrò che gli unici modelli di misurazione a essere caratterizzati dall'*oggettività specifica* sono quelli che ammettono statistiche sufficienti per i parametri e dimostrò che l'unico modello probabilistico di analisi degli item ad ammettere statistiche sufficienti è il Modello di Rasch. Per tale motivo il *SLM* e le sue estensioni rivestono un ruolo sempre più importante nelle analisi che richiedono la misurazione di variabili latenti.
Le due generalizzazioni più famose del SLM riguardano:

- il caso di struttura di risposta politomica (con le varianti *Rating Scale Model*, *Partial Credit Model* e *Extended Logistic Model*);

- il caso *multifacet*, dove le entità da misurare sono più di due (in genere, oltre ai soggetti e agli item si misura anche la severità dei correttori e la difficoltà dei temi cui gli item appartengono).

2.2.1 Il Rating Scale Model (RSM)

Nel 1978 David Andrich (Andrich, 1978) sviluppò una soluzione per la funzione punteggio e i coefficienti di categoria per il modello generale di Rasch espresso dalla 2.27, esprimendo ϕ_x e κ_x in termini di "soglie" (τ_k), livelli che dividono il *continuum* su cui si misura la variabile latente in categorie ordinali adiacenti, e di "discriminanti" delle soglie α_k. Una soglia rappresenta un limite sulla scala di misurazione oltre il quale una categoria di risposta diventa più probabile di quella che la precede; se le categorie di risposta costituiscono una scala a m categorie, le soglie da stimare sono pari a m - *1.*
Se $x = 1, 2, ..., m$ è il punteggio realizzato in un item, ponendo

$$\phi_x = \alpha_1 + \alpha_2 + ... + \alpha_k + ... + \alpha_x \quad e$$

$$\kappa_x = -(\alpha_1\tau_1 + \alpha_2\tau_2 + ... + \alpha_k\tau_k + ... + \alpha_x\tau_x)$$

Andrich ottenne l'espressione del *RM* per più categorie di risposta, nel caso in cui tutti gli item abbiano lo stesso numero di categorie. In un secondo momento Andrich considerò i discriminanti delle soglie pari a 1, in modo tale che la funzione di punteggio divenne

$$\phi_x = x$$

e il vettore di costanti

$$\kappa_x = -\sum_{k=1}^{x} \tau_k \quad \text{con } \kappa_0 \equiv 0$$

ottenendo la formalizzazione del *Rating Scale Model* (*RSM*)

$$Pr\{X_{\nu i} = x; \theta_\nu, \delta_i, \tau_k\} = \frac{exp(x(\theta_\nu - \delta_i) - \sum_{k=1}^{x} \tau_k)}{\sum_{x=0}^{m} exp(x(\theta_\nu - \delta_i) - \sum_{k=1}^{x} \tau_k)}. \tag{2.28}$$

Il *RSM* viene utilizzato nei casi in cui tutti gli item hanno la stessa struttura e lo stesso numero di categorie di risposta, tipicamente nei questionari con domande che prevedono una scala di Likert fissa. Il parametro τ_k specializza

il grado di difficoltà medio dell'item (δ_i) in ciascuna classe che lo compone e i valori delle soglie devono essere ordinati in modo da rispettare l'ordine predefinito delle categorie di risposta; ad esempio se le categorie di risposta sono tre con i livelli 0, 1 e 2, dove lo zero rappresenta totale assenza, 1 una presenza incompleta e 2 presenza completa di una caratteristica latente da misurare, deve sussistere che $\tau_0 < \tau_1 < \tau_2$.

Sebbene il vettore delle soglie sia assunto costante per tutti gli item, nessun vincolo (se non quello di ordinamento) è imposto alle distanze tra le soglie; la distanza che intercorre tra due categorie adiacenti $\tau_h - \tau_{h-1}$ è indipendente dalla distanza di qualsiasi altra coppia di soglie adiacenti.

Il *RSM* conserva una proprietà fondamentale dei Modelli di Rasch: la *sufficienza* delle statistiche per la stima dei parametri. Il che equivale ad affermare che anche per il *RSM* vale la separabilità delle stime dei parametri θ_n e δ_i (Andrich, 1982b; Fischer e Molenaar, 1995, cap.16).

2.2.2 Il Partial Credit Model (PCM)

Master (Master, 1982) propose un'altra estensione del *RM*, il *Partial Credit Model* (*PCM*), considerando la possibilità che le categorie di risposta potessero essere diverse da item ad item. Egli ipotizzò che tra due categorie di risposta adiacenti esistesse uno *step* di difficoltà che il soggetto doveva superare per poter raggiungere la categoria di risposta superiore.
Il modello dicotomico di Rasch si può esprimere come:

$$\phi_{\nu i1} = \frac{exp\{\theta_\nu - \delta_{i1}\}}{1 + exp\{\theta_\nu - \delta_{i1}\}} \tag{2.29}$$

dove $\phi_{\nu i1}$ è la probabilità che la persona ν realizzi 1 piuttosto che 0 nell'item i, θ_ν è la sua abilità e δ_{i1} è la difficoltà dello *step* nell'item i. L'equazione 2.29 specifica, inoltre, il modo in cui la probabilità di successo nell'item i è governata dall'abilità della persona e dalla difficoltà dell'item. Per lo sviluppo del modello è utile introdurre $\pi_{\nu i0}$, la probabilità che la persona ν ottenga 0 nell'item i, e $\pi_{\nu i1}$, la probabilità che ottenga 1. Nel caso dicotomico questa nuova notazione è ridondante in quanto, essendoci un solo *step* (da 0 a 1),

risulta $\pi_{\nu i1} = \phi_{\nu i1}$ e $\pi_{\nu i0}$ = 1 - $\pi_{\nu i1}$. La

$$\phi_{\nu i1} = \frac{\pi_{\nu i1}}{\pi_{\nu i0} + \pi_{\nu i1}} = \frac{exp\{\theta_\nu - \delta_{i1}\}}{1 + exp\{\theta_\nu - \delta_{i1}\}} \tag{2.30}$$

esplicita il fatto che $\phi_{\nu i1}$ è la probabilità che l'individuo ν ottenga 1 anziché 0 nell'item i.

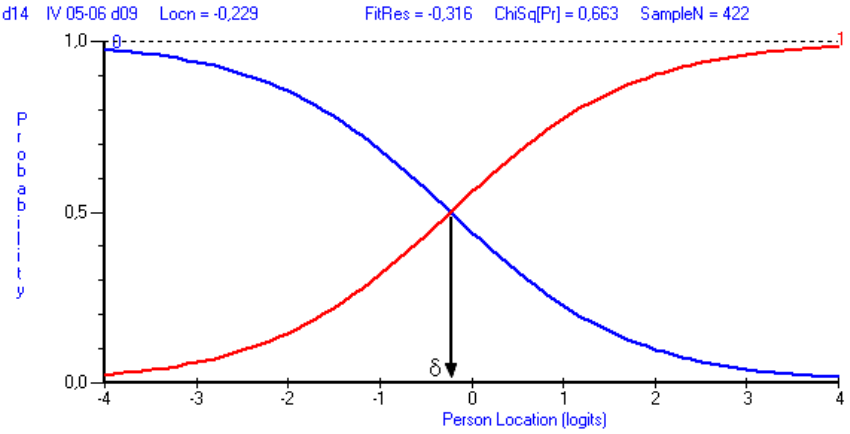


Figura 2.1: Curve di probabilità nel caso dicotomico

La figura 2.1 indica le due curve di probabilità del modello dicotomico, cioè come variano $\pi_{\nu i0}$ e $\pi_{\nu i1}$ al variare dell'abilità.

Il *PCM* suppone che nel rispondere all'item ci siano dei livelli intermedi di difficoltà crescente, a cui assegnare dei punteggi intermedi, tra la riposta totalmente sbagliata e la risposta totalmente corretta. Si consideri il caso di un item con due *step*, con livelli di performance 0, 1 e 2. L'espressione per la probabilità che l'individuo ν ottenga 1 anziché 0 è identica a quella del caso dicotomico della formula 2.30. L'unica differenza è che in questo caso, poiché si considerano più di due livelli di performance, $\pi_{\nu i0} + \pi_{\nu i1} < 1$; δ_{i1} governa ancora la probabilità di completare il primo *step* e ottenere 1 anzichè 0, ma il primo *step* non è più l'unico *step*. La 2.30 diventa una probabilità condizionata, date due possibili alternative: scegliere la categoria 0 o scegliere la categoria 1. Il secondo *step*, dal livello 1 al livello 2, si può raggiungere solo avendo superato il primo *step*, cioè dopo essere passati dal livello 0 al livello 1. L'espressione per la probabilità di completare questo secondo *step* nell'item i

è data da:

$$\phi_{\nu i2} = \frac{\pi_{\nu i2}}{\pi_{\nu i1} + \pi_{\nu i2}} = \frac{exp\{\theta_\nu - \delta_{i2}\}}{1 + exp\{\theta_\nu - \delta_{i2}\}} \tag{2.31}$$

che esprime la probabilità che l'individuo ν ottenga 2 anziché 1 nell'item i, come funzione dell'abilità θ_ν e di un secondo parametro δ_{i2} che governa la probabilità di completare lo *step* dal livello 1 al livello 2. Se da un lato δ_{i2} governa la probabilità di completare lo *step* dal livello 1 al livello 2, dall'altro esso nulla dice circa la probabilità di raggiungere il livello 1, che dipende ancora una volta dall'abilità θ_ν e dalla difficoltà del primo *step* nell'item δ_{i1}.

Per item con più di due *step*, si possono ricavare delle espressioni di probabilità analoghe alle 2.30 e 2.31. In generale, se si organizzano le categorie di risposta adiacenti in coppie dicotomiche, ciascuna con una relazione d'ordine tale che $k-1 < k$, e si assegna un parametro di difficoltà δ_{ik}, specifico per ciascun item, che regola il passaggio dal livello $k-1$ al livello k, assumendo che la probabilità condizionata di scegliere la soglia superiore aumenti all'aumentare dell'abilità θ_ν, il *PCM* si può esprimere mediante la

$$\phi_{\nu ik} = \frac{\pi_{\nu ik}}{\pi_{\nu ik-1} + \pi_{\nu ik}} = \frac{exp\{\theta_\nu - \delta_{ik}\}}{1 + exp\{\theta_\nu - \delta_{ik}\}} \quad k = 1, 2, ..., m_i \tag{2.32}$$

che viene utilizzata per descrivere la probabilità di rispondere ad ogni sequenza ordinata di *step* dicotomici.
Come nel caso di un item dicotomico in cui l'*ICC* rappresenta la probabilità di rispondere 1 anziché 0, così in presenza di item con più *step* è possibile rappresentare su un piano cartesiano le ogive, ognuna delle quali rappresenta, per lo stesso item, la probabilità di rispondere al k-esimo *step*, k anzichè *k-1.*

La formalizzazione del *PCM*, in termini di probabilità assolute, è data da

$$\pi_{\nu ix} = \frac{exp\sum_{j=0}^{x}(\theta_\nu - \delta_{ij})}{\sum_{k=0}^{m_i} exp\sum_{j=0}^{k}(\theta_\nu - \delta_{ij})} \quad x = 0, 1, ..., m_i \tag{2.33}$$

dove $\delta_{i0} \equiv 0$ in modo che $\sum_{j=0}^{0}(\theta_\nu - \delta_{ij}) = 0$, $exp\sum_{j=0}^{0}(\theta_\nu - \delta_{ij}) = 1$ e $\sum_{j=0}^{k}(\theta_\nu - \delta_{ij}) \equiv \sum_{j=1}^{k}(\theta_\nu - \delta_{ij})$.

Nella 2.33 la x indica il conteggio degli *step* completati; il numeratore contiene solamente le difficoltà δ_{i1}, δ_{i2}, ..., δ_{ix} degli x *step* completati mentre

il denominatore, che funge da fattore di normalizzazione, è la somma di tutti i possibili $m_i + 1$ numeratori. Per la stima dei parametri δ_{ij} è necessario specificare un'origine arbitraria sulla scala; generalmente ciò avviene ponendo il vincolo che la difficoltà media delle soglie sia nulla.

Analogamente a quanto avviene per il *RSM*, non viene formulata nessuna assunzione circa la distanza tra le categorie e, come nel *RSM*, valgono le proprietà di separabilità delle stime dei parametri dei soggetti dalle stime delle difficoltà delle soglie degli item (per cui anche in questo modello sussistono i presupposti per una misurazione oggettiva dei parametri) e le quantità θ_ν e δ_{ij} ammettono statistiche sufficienti per la loro stima. Per l'abilità θ_ν la statistica è

$$r_\nu = \sum_{i=1}^{I} x_{\nu i}$$

cioè il punteggio ottenuto dall'individuo ν negli I item, mentre il parametro δ_{ij} è stimato da

$$s_{ij} = \sum_{\nu=1}^{N} S_{\nu ij}$$

cioè dal numero di soggetti che superano la j-esima soglia di risposta dell'item i.

Come nel caso dicotomico δ_{i1} si trova all'intersezione delle curve di probabilità relative alle categorie 0 e 1, il secondo parametro δ_{i2}, si trova all'intersezione delle curve di probabilità relative alle categorie 1 e 2. Se il primo *step* dell'item i è più facile e il secondo *step* più difficile, il parametro δ_{i1} si trova a sinistra di δ_{i2}, ben distanziato da questi, e la curva di probabilità per la categoria di risposta intermedia, $\pi_{\nu i1}$, è molto alta, il che sta a significare che la probabilità di completare solo il primo *step* è maggiore per ogni valore di θ_ν. Se, al contrario il secondo *step* è più facile del primo, la probabilità $\pi_{\nu i1}$ si riduce, indicando che la *chance* di superare solo il primo *step* è molto bassa, e i parametri δ_{i1} e δ_{i2} risultano molto ravvicinati (al limite il loro ordine sulla scala si inverte). Un caso del genere indica un malfunzionamento dell'item nel misurare l'abilità latente. Infatti il RM presuppone che a punteggi più alti corrispondano abilità maggiori, il che non avviene se il secondo *step* è più facile

del primo.

La 2.33 esprime il *PCM* in funzione dell'abilità latente θ_ν dell'individuo e dei parametri degli *step* δ_{i1}, δ_{i2}, ..., δ_{im_i}. È possibile esprimere lo stesso modello utilizzando un altro *set* di parametri, γ_{i1}, γ_{i2}, ..., γ_{im_i}, le cosiddette soglie (*category boundaries*), che rappresentano i livelli ordinati di difficoltà delle categorie. Se si sostituisce lo *score* $x_{\nu i}$ dell'individuo con una colonna di *score* dicotomici $y_{\nu ik}$ per ogni livello di performance, per analizzare i dati è possibile utilizzare il RM dicotomico nella forma:

$$\pi^*_{\nu ik} = Pr\left\{y_{\nu ik} = 1; \theta_\nu, \gamma_{ik}\right\} = \frac{exp(\beta_\nu - \gamma_{ik})}{1 + exp(\beta_\nu - \gamma_{ik})} \tag{2.34}$$

dove θ_ν è l'abilità della persona e γ_{ik} è la difficoltà di raggiungere il livello k nell'item i (Wright, Masters; 1982).

Una delle assunzioni principali su cui si basa il *SLM* è che ogni osservazione $y_{\nu ik}$ sia determinata, a meno di errori casuali, esclusivamente da un parametro dell'individuo e un parametro dell'item. Ma in questo caso è impossibile raggiungere il livello k se prima non è stato raggiunto il livello $k-1$, e prima ancora il livello $k-2$, e così via. Quindi, che l'individuo raggiunga o meno il livello k dipende non solo dalla difficoltà del k-esimo *step*, ma anche dalla difficoltà di tutti gli *step* precedenti. Tale dipendenza gerarchica contrasta pertanto con l'assunto che la $Pr\left\{y_{\nu ik} = 1\right\}$ sia funzione solamente di θ_ν e γ_{ik}.

È possibile determinare la probabilità che l'individuo S_ν realizzi x nell'item I_i a partire dalla 2.34, seguendo un procedimento che sottrae, passo dopo passo, le probabilità cumulate. Nel caso di tre categorie di risposta e due *step* le probabilità sono:

$$\pi_{\nu i0} = 1 - \pi^*_{\nu i1} = \frac{1 + exp(\beta_\nu - \gamma_{i2})}{\Psi} \tag{2.35}$$

$$\pi_{\nu i1} = \pi^*_{\nu i1} - \pi^*_{\nu i2} = \frac{exp(\theta_\nu - \gamma_{i1}) + exp(\theta_\nu - \gamma_{i2})}{\Psi} \tag{2.36}$$

$$\pi_{\nu i2} = \pi^*_{\nu i2} - \pi^*_{\nu i3} = \frac{exp(\theta_\nu - \gamma_{i2}) + exp(2\theta_\nu - \gamma_{i1} - \gamma_{i2})}{\Psi} \tag{2.37}$$

con Ψ che rappresenta la somma dei tre numeratori.

2.2.3 Il Rasch's Extended Logistic Model (*ELM*)

Qualche anno più tardi Andrich (1985, 1988b) propose un nuovo modello, noto come *Extended Logistic Model* (*ELM*), con cui superò i limiti del *RSM*. Nell'*ELM* si assume che ad ogni item sia associato un vettore di valori di soglia (*category boundaries*) che rappresentano i livelli ordinati delle categorie dell'item. Per rappresentare il caso più generale le distanze tra soglie adiacenti sono lasciate libere di variare. La formulazione dell'*ELM* si presenta come segue:

$$Pr\{X_{\nu i} = x; \delta_i, \theta_\nu; \tau_{ki}\} = \frac{exp(x(\theta_\nu - \delta_i) - \sum_{k=1}^{x} \tau_{ki})}{\sum_{x=0}^{m_i} exp(x(\theta_\nu - \delta_i) - \sum_{k=1}^{x} \tau_{ki})}. \tag{2.38}$$

I parametri del modello *ELM* hanno un significato analogo a quelli del *RSM* nella 2.28, con l'unica differenza che qui le soglie sono indicizzate con $i = 1, 2, ..., I$ per indicare che ciascun item possiede un insieme specifico di soglie che separano le $m_i + 1$ categorie ordinate (dove m_i è il punteggio massimo nell'item i).

2.2.4 Il Multifacet Model

Il Modello Multifacet (*MM*) considera una struttura di riferimento in cui le "entità" che si incontrano, e si misurano, non sono più due (abilità degli individui e difficoltà degli item), ma diventano tre, quattro o più. Per ogni nuova "entità" che interviene nel processo di misurazione viene aggiunto al modello un parametro da stimare (sempre rispettando la condizione di additività).
Il caso più comune contempla un terzo soggetto (il giudice) che corregge le risposte e dal cui giudizio dipende parte dell'esito della prova. L'espressione della probabilità, nel caso di item dicotomico, è:

$$Pr\{X_{\nu ij} = 1; \theta_\nu, \delta_i, \gamma_j\} = \frac{exp(\theta_\nu - \delta_i - \gamma_j)}{1 + exp(\theta_\nu - \delta_i - \gamma_j)} \tag{2.39}$$

dove θ_ν è l'abilità dell'individuo S_ν, δ_i la difficoltà dell'item I_i e γ_j la severità del giudice P_j. Se, invece, i giudizi sono espressi su una scala di Likert, la probabilità è tale che:

$$Pr\{X_{\nu ij}\} : ln\frac{Pr\{X_{\nu ij} = k\}}{Pr\{X_{\nu ij} = k-1\}} = \theta_\nu - \delta_i - \gamma_j - \tau_k \qquad k = 0,\ 1,\ 2,\ \ldots,\ K \tag{2.40}$$

estensione del *RSM* espresso dalla 2.28, dove τ_k è il parametro della "soglia" che separa la *(k-1)*-esima dalla k-esima categoria.
Ammettendo la possibilità che ogni item abbia il proprio set di "soglie" si ottiene un'estensione del *PCM* in presenza di un giudice:

$$Pr\{X_{\nu ij}\} : ln\frac{Pr\{X_{\nu ij} = k\}}{Pr\{X_{\nu ij} = k-1\}} = \theta_\nu - \delta_i - \gamma_j - \tau_{ik} \qquad k = 0,\ 1,\ 2,\ \ldots,\ K. \tag{2.41}$$

Seguendo Lynch e McNamara (1998) è possibile valutare lo scostamento dalle ipotesi di linearità del RM e quindi l'eventuale distorsione di ciascun giudice, rispetto alle prove e/o ai soggetti. A tal fine, viene inserito nel modello un termine d'interazione:

$$Pr\{X_{\nu ij}\} : ln\frac{Pr\{X_{\nu ij} = k\}}{Pr\{X_{\nu ij} = k-1\}} = \theta_\nu - \delta_i - \gamma_j - \tau_{ik} + C_{\nu j} \tag{2.42}$$

nel caso si postuli un'interazione tra soggetti e giudici che consente di valutare se un giudice si sta comportando nello stesso modo rispetto a tutti i soggetti ($C_{\nu j} = 0 \quad \forall \nu$) o, piuttosto, non abbia assegnato, rispetto al suo stile di valutazione, punteggi troppo bassi o troppo alti a qualche soggetto; oppure

$$Pr\{X_{\nu ij}\} : ln\frac{Pr\{X_{\nu ij} = k\}}{Pr\{X_{\nu ij} = k-1\}} = \theta_\nu - \delta_i - \gamma_j - \tau_{ik} + C_{ij} \tag{2.43}$$

nel caso si postuli un'interazione tra prove e giudici che consente di valutare se un giudice mostra lo stesso comportamento rispetto a tutte le prove ($C_{ij} = 0 \quad \forall i$) o, piuttosto, non si comporti in maniera differente nel caso di qualche prova.

2.3 Violazioni del modello

Le principali violazioni dei dati al *SLM* sono il *differential item functioning* (*DIF*), la multidimensionalità (dipendenza di costrutto), la dipendenza locale (dipendenza di risposta), il *cheating* e il *guessing*. Tutte queste violazioni inficiano la proprietà dell'invarianza delle stime dei parametri rispetto alla popolazione considerata e viziano, se non addirittura impediscono, la costruzione della scala di misurazione della variabile latente. In questo paragrafo si descrivono il *DIF*, la dipendenza di costrutto e la dipendenza di risposta.

2.3.1 Differential Item Functioning

In particolare il *DIF*, conosciuto in letteratura anche con il nome di *item bias*, si manifesta quando, condizionatamente ad un livello di abilità, la probabilità di rispondere correttamente ad un item è diversa tra gruppi di individui omogenei secondo una certa caratteristica: un item è considerato *biased* solo se la sua difficoltà cambia tra individui con lo stesso livello di abilità, ma che appartengono a sottopopolazioni distinte.

Nella Item Response Theory, di cui i modelli di Rasch fanno parte, generalmente si fa una distinzione tra *DIF* uniforme (fig. 2.2) e *DIF* non uniforme (fig. 2.3) . Nel *DIF* uniforme la probabilità di una risposta corretta di una sottopopolazione è sistematicamente maggiore o minore della probabilità di una risposta corretta di un'altra sottopopolazione (in genere la sottopopolazione di riferimento) per tutti i livelli di abilità; nel *DIF* non uniforme la probabilità di una risposta corretta nella sottopopolazione di studio è maggiore della probabilità di risposta corretta della popolazione di riferimento nei livelli più bassi di abilità e minore nei livelli più alti, o viceversa.

Con riferimento alle Item Characteristic Curve (ICC) possiamo affermare che: i) le posizioni delle curve (quindi le difficoltà degli item) sono differenti ma la loro pendenza è la stessa; ii) le posizioni sono le stesse ma le pendenze sono diverse; iii) sia le posizioni che le pendenze sono diverse.

Tra i numerosi metodi statistici proposti per testare la presenza di *DIF* in una batteria di item, l'Analisi della Varianza dei residui standardizzati è

quello che ha trovato maggior adesione tra gli studiosi del RM, sia per la sua semplicità concettuale e applicativa sia per l'attendibilità dei risultati. L'Analisi della Varianza standard (ANOVA) dei residui standardizzati, scompone la varianza totale dei residui standardizzati

$$z_{n_{cg}i} = \frac{x_{n_{cg}i} - E(X_{n_{cg}i})}{\sqrt{var(X_{n_{cg}i})}} \tag{2.44}$$

dove c è l'indice della classe intervallo (CI) d'appartenenza, g il livello del fattore possibile causa del DIF, per testare la significatività dell'effetto principale del fattore, dell'effetto della CI e dell'effetto prodotto dall'interazione tra i due. Sebbene i residui standardizzati non siano lineari e solo approssimativamente normalmente distribuiti, l'affidabilità della distribuzione F usata nell'ANOVA produce una buona prova della presenza o meno del DIF e del buon funzionamento generale dell'item, analogamente a quanto prodotto dal test del chi-quadrato. Ogni valutazione statistica va comunque accompagnata dall'osservazione del grafico dell'ICC e da considerazioni metodologiche riguardanti la costruzione degli item.

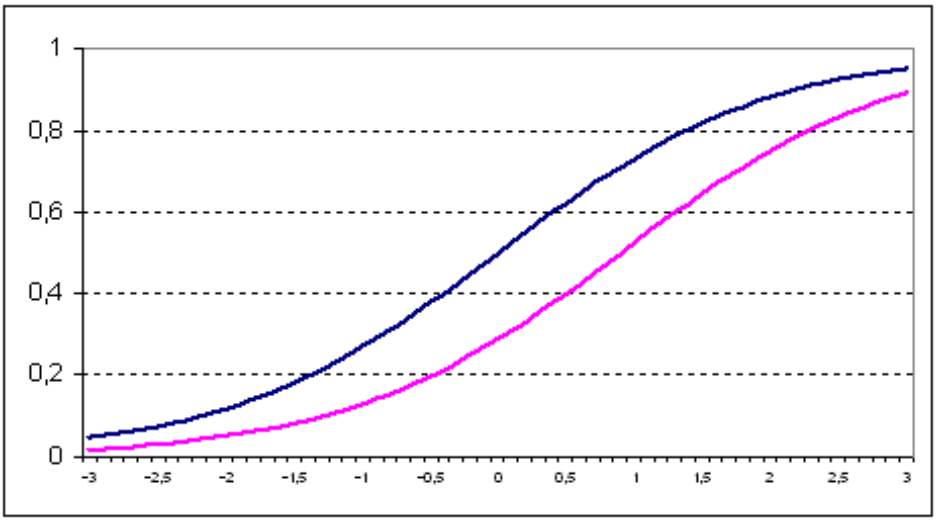


Figura 2.2: DIF uniforme

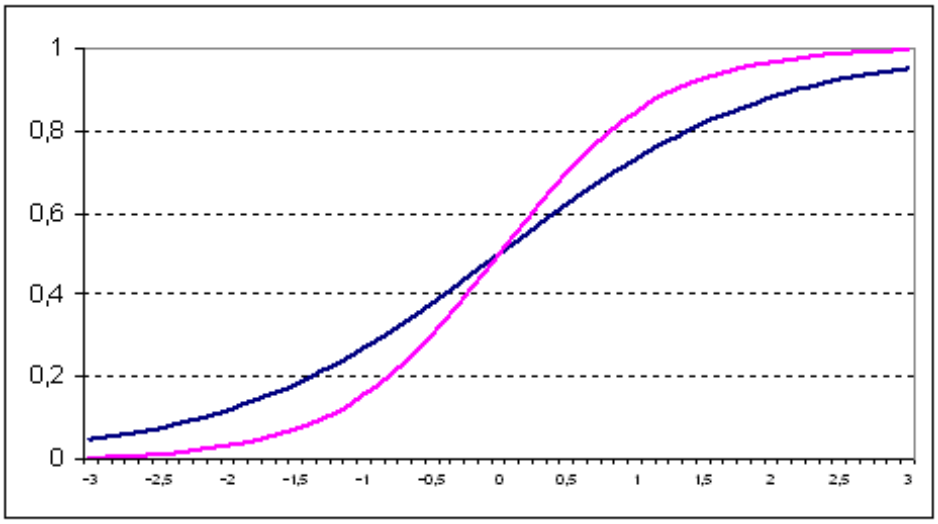


Figura 2.3: DIF non uniforme

2.3.2 Dipendenza di costrutto e dipendenza di risposta

L'ipotesi di indipendenza locale (*local independence*) nel RM può essere violata in due modi, che, generalmente, in letteratura, non sono ben distinti l'uno dall'altro; una violazione riguarda la dipendenza di costrutto (*trait dependence*) e l'altra la dipendenza di risposta (*response dependence*) (Marais, Andrich; 2008). L'indipendenza locale si esplica nel fatto che la relazione tra gli item è completamente spiegata dai soli parametri a essi riferiti, senza che ci sia la possibilità di intervento di altri fattori. Tale proprietà, in termini matematici, si traduce in:

$$Pr\left\{X_1, X_2, \ldots, X_k|\theta_\nu\right\} = \prod_{i=1}^{k} Pr\left\{X_i|\theta_\nu\right\}. \tag{2.45}$$

La *trait dependence* attiene alla multidimensionalità dei tratti latenti che possono emergere dal processo di misurazione, cioè quando le probabilità di risposta sono influenzate da altri parametri relativi ai soggetti, oltre che da θ_ν. È un caso che si riscontra abbastanza frequentemente nei test, composti da molti quesiti (spesso raggruppati in sottoinsiemi che hanno in comune la stessa struttura o lo stesso contenuto), che vogliono cogliere nelle persone

caratteristiche generali che, per loro natura, non costituiscono un costrutto unidimensionale.

Una formalizzazione algebrica della *trait dependence* considera appunto S sottogruppi di item in cui può essere suddiviso un test, ognuno dei quali coglie un tratto latente comune principale, θ_ν, e un tratto latente specifico del sottogruppo, $\theta'_{\nu s}$, che può essere considerato incorrelato con il primo ($cov[\theta, \theta'_s] = 0$). Pertanto con:

$$\theta_{\nu s} = \theta_\nu + c_s \theta'_{\nu s} \tag{2.46}$$

si rappresenta il parametro che governa le probabilità di risposta del soggetto ν nel sottogruppo s, dove $c_s > 0$ caratterizza la grandezza della variabile specifica del sottogruppo.

Senza perdere in generalità, si può pensare che le variabili dei sottogruppi siano mutualmente incorrelate, cioè $cov[\theta'_s, \theta'_t] = 0$ per ogni $s \neq t$. L'equazione del modello logistico diventa allora:

$$Pr\{X^s_{\nu i}; \theta_{\nu s}, \delta_i\} = \frac{exp\{X^s_{\nu i}(\theta_{\nu s} - \delta_i)\}}{1 + exp\{\theta_{\nu s} - \delta_i\}} \tag{2.47}$$

dove l'apice s indica il sottogruppo $s = 1, 2, \ldots, S$ di appartenenza dell'item. Tale equazione viola i requisiti del *SLM* espresso dalla 1.36. Inoltre, poiché ogni sottogruppo è composto da una variabile specifica $\theta'_{\nu s}$ e da una variabile comune θ_ν, la correlazione tra tratti latenti di sottogruppi distinti non è uguale a 1, ma, generalmente, assume un valore maggiore di 0 e minore di 1, dipendendo dal valore di c_s e c_t.
Infatti, imponendo la condizione che le varianze delle abilità sottese dal modello siano uguali, cioè che:

$$Var(\theta) = Var(\theta'_s) = Var(\theta'_t) = \sigma^2 \tag{2.48}$$

risulta che:

$$Var(\theta_s) = Var(\theta + c_s\theta'_s) = Var(\theta) + c_s^2 Var(\theta'_s) = (1 + c_s^2)\sigma^2,$$

$$Var(\theta_t) = Var(\theta + c_t\theta'_t) = Var(\theta) + c_t^2 Var(\theta'_t) = (1 + c_t^2)\sigma^2,$$

$$Cov(\theta, \theta_s^{'}) = Cov(\theta, \theta_t^{'}) = Cov(\theta_s^{'}, \theta_t^{'}) = 0,$$

$$Cov(\theta_s, \theta_t) = Cov(\theta + c_s\theta_s^{'}, \theta + c_t\theta_t^{'}) = Cov(\theta, \theta) = \sigma^2.$$

Allora si trova che la correlazione tra i tratti di ciascuna coppia di item appartenenti a due differenti sottogruppi s e t è:

$$\rho_{st} = \frac{Cov[\theta_s, \theta_t]}{\sqrt{Var(\theta_s)}\sqrt{Var(\theta_t)}} = \frac{1}{\sqrt{1+c_s^2}\sqrt{1+c_t^2}}. \tag{2.49}$$

Se, poi, $c_s = c_t = c$ la 2.49 si riduce a:

$$\rho_{st} = \frac{1}{\sqrt{1+c^2}}. \tag{2.50}$$

Per K_s item nel sottoinsieme s e K_t item nel sottoinsieme t, i punteggi latenti per l'individuo ν sono uguali a:

$$K_s\theta_{\nu s} = K_s\theta_\nu + K_s c_s\theta_{\nu s}^{'} \quad \text{e} \quad K_t\theta_{\nu t} = K_t\theta_\nu + K_t c_t\theta_{\nu t}^{'} \tag{2.51}$$

con

$$Var(K_s\theta_s) = K_s^2 Var(\theta) + K_s^2 c_s^2 Var(\theta_s^{'}),$$

$$Var(K_t\theta_t) = K_t^2 Var(\theta) + K_t^2 c_t^2 Var(\theta_t^{'}) \quad \text{e}$$

$$Cov(K_s\theta_s, K_t\theta_t) = K_s K_t Var(\theta).$$

Ne segue che la correlazione tra due differenti sottoinsiemi è pari alla correlazione tra i tratti latenti di ciascuna coppia di item (I_s, I_t) con $s \neq t$, cioè:

$$\rho_{st}^{'} = \frac{1}{\sqrt{1+c_s^2}\sqrt{1+c_t^2}} = \rho_{st}. \tag{2.52}$$

La *response dependence*, che per alcuni studiosi costituisce la dipendenza locale vera e propria, attiene al fatto che, per uno stesso individuo, la risposta a un item è influenzata dalla risposta data a un altro item. Si riscontra nei test dove lo svolgimento o la soluzione di alcune domande fornisce indicazioni

su come rispondere ad altre domande, oppure alcuni quesiti sono molto simili tra loro, o, ancora, la risposta ad alcuni item è necessaria per la soluzione di altri item.

Una formalizzazione algebrica della *response dependence* si ricava inserendo nel modello *SLM* una costante $d > 0$ che, sottratta o sommata, rispettivamente, alla difficoltà dell'item indipendente I_i, definisce il livello di difficoltà dell'item dipendente I_j, ottenendo:

$$Pr\{X_{\nu j} = 1 | X_{\nu i} = 1\} = \frac{exp[\theta_\nu - (\delta_i - d)]}{1 + exp[\theta_\nu - (\delta_i - d)]} \quad \text{e} \tag{2.53}$$

$$Pr\{X_{\nu j} = 1 | X_{\nu i} = 0\} = \frac{exp[\theta_\nu - (\delta_i + d)]}{1 + exp[\theta_\nu - (\delta_i + d)]}. \tag{2.54}$$

Le 2.53 e 2.54 possono essere sintetizzate in un'unica espressione come segue:

$$Pr\{X_{\nu j} = 1 | X_{\nu i} = x_i\} = \frac{exp[\theta_\nu - \delta_i - (1 - 2x_i)d]}{1 + exp[\theta_\nu - \delta_i - (1 - 2x_i)d]}. \tag{2.55}$$

La generalizzazione che considera entrambe le possibilità di risposta per l'item I_j è:

$$Pr\{X_{\nu j} = x_j | X_{\nu i} = x_i\} = \frac{exp[x_j(\theta_\nu - \delta_i - (1 - 2x_i)d)]}{1 + exp[\theta_\nu - \delta_i - (1 - 2x_i)d]}. \tag{2.56}$$

Studi di simulazione (Marais e Andrich, 2008) evidenziano che violazioni dell'indipendenza locale modificano l'unità della scala di misura, manifestandosi in variazioni del *range* e della *standard deviation* degli item e della *standard deviation* degli individui, con effetti opposti a seconda che si tratti di *trait dependence* o *response dependence*; nel caso di *trait dependence* la scala si riduce mentre nel caso di *response dependence* la scala si espande. Le variazioni di scala si ripercuotono anche sull'indice di separazione che diminuisce in presenza di *trait dependence* e aumenta in presenza di *response dependence* e creano distorsioni di un certo rilievo nelle stime delle abilità latenti se si costruisce una scala di misura mediante un *link* di più test (Humphry, 2005; Humphry, 2006).

I principali strumenti per rilevare violazioni dell'unidimensionalità dei dati sono due: il *fit* degli item e l'analisi delle componenti principali dei residui standardizzati (Smith R., 1992 e 1996; Smith e Miao, 1994). Studi di simulazione (Smith E. V., 2002) hanno evidenziato che, in caso di multidimensionalità, quando le componenti che definiscono i diversi tratti latenti hanno un numero quasi uguale di item e le componenti non sono molto correlate, l'analisi delle componenti principali dà risultati migliori; invece, quando le componenti sono molto correlate e la maggior parte degli item contribuisce a definire una sola componente, l*Infit* e l'*Outfit* standardizzati funzionano meglio.

Il *fit* degli item e l'analisi delle *ICC* sono largamente utilizzati per verificare la presenza di *response dependence*. Un metodo alternativo è stato proposto da Andrich (2009), che fornisce una procedura sia per stimare il parametro d delle 2.53, 2.54, 2.55 e 2.56, sia per ridurre l'effetto della dipendenza tra item in un test.

Se il set di risposte $\{X_{\nu j} = x_j\}$ all'item dipendente I_j viene diviso in due item, uno, I_{j1}, relativo alle persone che hanno risposto in modo corretto all'item indipendente I_i, e uno, I_{j0}, relativo alle persone che hanno risposto in modo errato all'item I_i, ponendo valore mancante nei casi non pertinenti, i due nuovi item non si sovrappongono mai e sono, pertanto, indipendenti. Però, poiché sono entrambi dipendenti da I_i, essi sono indirettamente dipendenti; eliminando l'item I_i dal test la dipendenza scompare.
Si supponga che gli item soddisfino il RM e che il numero di item del test sia sufficiente a stimare le difficoltà di I_{j0} e I_{j1}, $\hat{\delta}_{ji0}$ e $\hat{\delta}_{ji1}$. Ne consegue che:

$$\hat{\delta}_{ji0} = \hat{\delta}_j + \hat{d} \quad \text{e} \quad \hat{\delta}_{ji1} = \hat{\delta}_j - \hat{d}$$

da cui:

$$\hat{d} = \frac{\hat{\delta}_{ji0} + \hat{\delta}_{ji1}}{2}$$

con *standard error*

$$\hat{SE}_d = \sqrt{SE^2_{ji0} + SE^2_{ji1}}$$

da cui si può testare la significatività di $d \neq 0$.

Andrich, inoltre, fa vedere che la dipendenza locale a livello di item dico-

tomici scompare o si riduce di molto se gli item che violano l'indipendenza vengono aggregati per formare item politomici; la violazione di indipendenza viene assorbita dai parametri delle soglie che mostrano un malfunzionamento.

2.4 L'Analisi di adattamento dei dati al modello

L'analisi statistica di adattamento dei dati al modello (analisi del fit) è di primaria importanza per l'interpretazione e l'attendibilità delle stime; essa fornisce una misura di quanto bene le risposte osservate si avvicinano ai valori attesi del modello e con quanta plausibilità le risposte di ogni individuo sono attendibili valendo le ipotesi del modello considerato.

Ci sono tre aspetti da considerare quando si parla dell'analisi del *fit*:

1. le assunzioni e le proprietà del modello da testare;
2. il tipo di statistiche utilizzato per il test;
3. la matematica della procedura.

Il primo aspetto si riferisce alle sufficienza delle statistiche degli item e degli individui, alla monotonicità e parallelismo delle *ICC*, all'unidimensionalità, all'indipendenza locale e al *Differential Item Functioning* (*DIF*), il secondo alle procedure adottate per testare le violazioni alle assunzioni del modello, (statistiche basate sul test di Pearson, che confrontano frequenze osservate e frequenze attese, statistiche basate sul rapporto della verosimiglianza e statistiche basate sul test di Wald), il terzo agli strumenti statistico-matematici veri e propri utilizzati. In generale, ci sono due approcci per condurre l'analisi del fit nel Modello di Rasch; con il primo si stimano i parametri del modello sull'intero set dei dati e poi si controlla quanto bene i valori stimati dal modello siano in grado di riprodurre l'intera matrice osservata o determinate partizioni di essa; il secondo si basa sul principio, intrinseco nel RM, che i parametri stimati devono essere invarianti nelle varie partizioni dei dati, e tende a verificare che tale invarianza venga rispettata, confrontando le stime ottenute in ciascun sottogruppo del campione o della popolazione. Va sottolineato

che accade di frequente che una o più assunzioni del modello (soprattutto del *SLM*) vengano violate. Ciò è dovuto alla "rigidità" del modello e al suo limitato numero di parametri che non consente di prendere in considerazione una molteplicità di aspetti del fenomeno che intervengono nel determinare la matrice delle osservazioni.

2.4.1 Test di adattamento basati sui punteggi totali

Tra le varie procedure per testare l'adattamento dei dati al modello secondo l'approccio basato sui *total scores* (Andrich 1988a e 1988b, Bond e Fox 2007, Cristante e Mannarini 2004, Wright e Masters 1982) le più diffuse sono quelle che prendono in considerazione i residui, cioè le differenze tra le risposte di ciascun individuo a ciascun item e i loro valori attesi. Tutte utilizzano il test di adattamento del chi-quadrato , differendo tra loro per il modo in cui la statistica chi-quadrato viene costruita. Ogni risposta $X_{\nu i}$ viene confrontata con

$$E(X_{\nu i}) = \hat{P}_{\nu i}(X_{\nu i} = 1; \theta_\nu, \delta_i) = \frac{exp\{\hat{\theta}_\nu - \hat{\delta}_i\}}{1 + exp\{\hat{\theta_\nu} - \hat{\delta}_i\}} \tag{2.57}$$

in cui la probabilità $\hat{P_{\nu i}}$ è ottenuta sostituendo i parametri incogniti con le stime $\hat{\theta_\nu}$ - $\hat{\delta}_i$.
Il vincolo del processo di stima impone che i valori attesi di $X_{\nu i}$, sommati per riga e per colonna, debbano risultare pari agli *score* osservati, rispettivamente degli individui e degli item, cioè

$$E(X_{\nu 1}) + E(X_{\nu 2}) + ... + E(X_{\nu I}) = r_\nu \tag{2.58}$$

e

$$E(X_{1i}) + E(X_{2i}) + ... + E(X_{ni}) = s_i. \tag{2.59}$$

Ora, volendo quantificare quanto bene i totali, di riga e di colonna, riproducono la matrice delle osservazioni, si confronta ogni valore osservato $\tilde{x}_{\nu i}$ con il valore atteso $E(X_{\nu i})$ e si determina lo *score residual*:

$$Y_{\nu i} = \tilde{x}_{\nu i} - E(X_{\nu i})$$

$$= \tilde{x}_{\nu i} - \hat{P}_{\nu i}(X_{\nu i} = 1; \theta_\nu, \delta_i)$$

$$= \tilde{x}_{\nu i} - \frac{exp\{\hat{\theta}_\nu - \hat{\delta}_i\}}{1 + exp\{\hat{\theta}_\nu - \hat{\delta}_i\}} \tag{2.60}$$

Nel caso dicotomico lo *score residual* assume sempre uno dei due seguenti valori:

$$1 - \frac{exp\{\hat{\theta}_\nu - \hat{\delta}_i\}}{1 + exp\{\hat{\theta}_\nu - \hat{\delta}_i\}} \tag{2.61}$$

nel caso l'individuo S_ν abbia risposto correttamente all'item i e

$$-\frac{exp\{\hat{\theta}_\nu - \hat{\delta}_i\}}{1 + exp\{\hat{\theta}_\nu - \hat{\delta}_i\}} \tag{2.62}$$

nel caso l'individuo S_ν abbia risposto in modo errato all'item i.
Poichè, per i valori stimati, vale sempre:

$$0 < \frac{exp\{\hat{\theta}_\nu - \hat{\delta}_i\}}{1 + exp\{\hat{\theta}_\nu - \hat{\delta}_i\}} < 1$$

ne consegue che $-1 < Y_{\nu i} < 1$.

Il valore atteso dello *score residual* è sempre pari a 0:

$$E(Y_{\nu i}) = 0,$$

mentre la sua varianza:

$$Var(Y_{\nu i}) = \hat{P}_{\nu i}(1 - \hat{P}_{\nu i})$$

dipende dalle stime delle abilità delle persone e delle difficoltà degli item. Inoltre, residui negativi corrispondono sempre a risposte errate e residui positivi corrispondono sempre a risposte corrette e la loro somma, considerata per riga (relativa all'individuo S_ν) o per colonna (relativa all'item i) nella matrice dei dati, è sempre 0, cioè:

$$Y_{\nu 1} + Y_{\nu 2} + ... + Y_{\nu k} = 0 \quad \text{e}$$

$$Y_{1i} + Y_{2i} + ... + Y_{ni} = 0$$

Per correggere l'effetto causato della differenza di varianza degli $Y_{\nu i}$ si considera il residuo standardizzato (Wright, 1977):

$$Z_{\nu i} = \frac{Y_{\nu i}}{[\hat{P}_{\nu i}(1 - \hat{P}_{\nu i})]^{1/2}} \tag{2.63}$$

cioè si divide lo *score residual* per la sua *standard deviation* stimata.

I primi a utilizzare i residui standardizzati per valutare il fit degli item al RM furono Wright e Panchapakesan (1969) che, basandosi sui gruppi di *score* omogenei per item dicotomici, proposero un test costruito su residui standardizzati che aveva la forma:

$$R_{ij} = \frac{a_{ij} - r_j p_{ij}}{[r_j p_{ij}(1 - p_{ij})]^{1/2}} \tag{2.64}$$

dove a_{ij} rappresenta il numero osservato di risposte corrette all'item i delle persone che hanno ottenuto uno *score* pari a j, r_j è il numero di persone che hanno ottenuto lo *score* j e p_{ij} è la probabilità di rispondere correttamente all'item i per il gruppo di punteggio j.

Quasi la totalità dei test oggi utilizzati si basano sui residui standardizzati $Z_{\nu i}$, le cui principali proprietà sono:

1. valori positivi rappresentano sempre risposte corrette;

2. valori negativi rappresentano sempre risposte errate;

3. $Z_{\nu i}$ solitamente varia tra -10 e 10;

4. $E(Z_{\nu i}) = 0$ e $Var(Z_{\nu i}) = 1$;

5. $Z_{\nu i}$ è approssimativamente distribuito come una normale standard.

Residui standard più grandi di 2 si verificano solo quando la probabilità di una risposta corretta è inferiore a 0.2, mentre residui standard più piccoli di -2 si verificano solo quando la probabilità di una risposta corretta è superiore a 0.8; residui standard più grandi di 3 si verificano solo quando la probabilità di una

risposta corretta è inferiore a 0.1, mentre residui standard più piccoli di -3 si verificano solo quando la probabilità di una risposta corretta è superiore a 0.9.

Nonostante siano state sollevate alcune critiche riguardo l'utilizzo degli *standardized residual* nell'analisi del *fit*, gli studi di simulazione condotti da Richard M. Smith (1988) hanno evidenziato che quando c'è un buon adattamento dei dati al modello la distribuzione degli *standardized residual*, sotto una varietà di condizioni (numerosità e distribuzione degli individui e degli item), ha una media e una deviazione standard prossime a quelle teoriche e che si possono definire delle percentuali di errore di I tipo da usare come riferimento per testare la bontà del modello. In caso di dati che violano una o più proprietà (multidimensionalità, differenti coefficienti di discriminazione delle ICC, *guessing*), la distribuzione degli *standardized residual* si modifica ma le variazioni sono generalmente troppo piccole ed annullano la potenza delle statistiche costruite per testare le violazioni del modello.

Per correggere l'effetto del segno che annulla il valore dei residui di segno opposto, e per ottenere un utile indice di *fit* sommando i residui tra gli individui e gli item, si considera lo *squared standardized residual* $Z^2_{\nu i}$. Le principali proprietà dello *squared standardized residual* sono:

1. $Z^2_{\nu i}$ può essere considerato come una variabile chi-quadrato con 1 grado di libertà (quindi $E(Z^2_{\nu i}) = 1$ e $Var(Z^2_{\nu i}) = 2$);

2. $Z^2_{\nu i}$ assume valori compresi tra 0 e 100;

3. la maggior parte dei valori varia tra 0 e 1.

C'è da fare attenzione però, che la variabile $Z^2_{\nu i}$ si distribuisce solo approssimativamente come una chi-quadrato con 1 gdl; $Z^2_{\nu i}$ sarebbe una perfetta chi-quadrato se tutte le seguenti condizioni fossero soddisfatte:

1. le $P_{\nu i}$ fossero note e non stimate;

2. la variabile $X_{\nu i}$ fosse continua e non discreta;

3. I dati fittano con il modello.

Lo *squared standardized residual* è la statistica utilizzata nei test principali per esaminare se i dati osservati rispecchiano le condizioni del RM. Sommando gli $Z^2_{\nu i}$ su tutta la matrice si testa l'adattamento dell'intera matrice di dati al modello:

$$\sum_{\nu=1}^{n}\sum_{i=1}^{k} Z^2_{\nu i} \tag{2.65}$$

approssima una chi-quadrato con $(n-1)(k-1)$ gradi di libertà. Sommando gli $Z^2_{\nu i}$ su una riga si ottiene un fit delle risposte di una persona (*person fit*) mentre sommandoli su una colonna si ottiene un fit delle risposte di un determinato item (*item fit*). Poichè $\sum_{\nu=1}^{n} Z^2_{\nu i} \rightarrow \chi^2_{n-1}$ e $\sum_{i=1}^{k} Z^2_{\nu i} \rightarrow \chi^2_{k-1}$, per ogni valore di n e di k c'è un differente valore critico, le statistiche del *person fit* e dell'*item fit* sono trasformate in quadrati medi (*mean square residual*), dividendo la variabile casuale per i suoi gradi di libertà. Così

$$MS(UT)_{\nu} = \frac{\sum_{i=1}^{k} Z^2_{\nu i}}{k-1} \tag{2.66}$$

è l'*observed mean square residual* per il ν-esimo individuo, mentre

$$MS(UT)_{i} = \frac{\sum_{\nu=1}^{n} Z^2_{\nu i}}{n-1} \tag{2.67}$$

è l'*observed mean square residual* per l'i-esimo item.

Altri autori propongono una versione del *mean square residual* leggermente diversa, dividendo $\sum_{i=1}^{k} Z^2_{\nu i}$ e $\sum_{\nu=1}^{n} Z^2_{\nu i}$ rispettivamente per k e n al fine di attenuare la distorsione dovuta al fatto che le stesse osservazioni $x_{\nu i}$ vengono utilizzate per stimare sia i parametri degli item e delle persone (e quindi $\hat{P}_{\nu i}$) che gli *score residual* (Smith-Schumacker-Bush, 1998).

Poichè il *mean square residual* ha il valor atteso unitario ma una varianza che dipende dalla numerosità campionaria (e degli *item* e degli individui), dalla sua distribuzione e dalle stime $\hat{P}_{\nu i}$, Wright e Stone (1979) propongono una trasformazione logaritmica del $MS(UT)_i$ nel tentativo di approssimare la statistica di adattamento ad una distribuzione normale con valori di riferimento

uguali per ogni *item*, considerando la

$$t_i = [\ln(MS(UT)_i) + MS(UT)_i - 1][\frac{n-1}{8}]^{1/2}. \tag{2.68}$$

Il *mean square residual* (*Unweighted Mean Square Statistic*), è una media aritmetica non ponderata dei quadrati dei residui; ad ogni residuo è assegnato lo stesso peso e pertanto è più sensibile agli *outliers* (da cui il nome con cui è più conosciuta, *OUTFIT*, cioè *Outlier-sensitive Fit Statistics*), cioè alle risposte corrette di persone con scarsa abilità ad item difficili o alle risposte sbagliate di persone molto capaci ad item facili. Per ovviare a tale inconveniente e ottenere una statistica più sensibile alle risposte inattese delle persone che hanno un'abilità stimata prossima alla difficoltà dell'item in esame (o alle risposte inattese agli item che hanno una difficoltà prossima all'abilità della persona in esame), si calcola la versione ponderata del *mean square residual*, prendendo come pesi le varianze dei residui, ossia le funzioni di informazione. La *Weighted Mean Square Statistic* relativa all'*item i* è espressa dalla:

$$MS(WT)_i = \frac{\sum_{\nu=1}^{n} \hat{P}_{\nu i}(1-\hat{P}_{\nu i})Z_{\nu i}^2}{\sum_{\nu=1}^{n} \hat{P}_{\nu i}(1-\hat{P}_{\nu i})} = \frac{\sum_{\nu=1}^{n}(X_{\nu i}-\hat{P}_{\nu i})^2}{\sum_{\nu=1}^{n} \hat{P}_{\nu i}(1-\hat{P}_{\nu i})} \tag{2.69}$$

conosciuta come *INFIT*, cioè *Information-weighted Fit Statistics*. Quando i dati si adattano bene al modello $MS(WT)_i$ ha un valore atteso unitario e una deviazione standard pari a

$$sd[MS(WT)_i] = \frac{[\sum_{\nu=1}^{n} w_{\nu i} - \sum_{\nu=1}^{n} w_{\nu i}^2]^{1/2}}{\sum_{\nu=1}^{n} w_{\nu i}} \tag{2.70}$$

con $w_{\nu i} = \hat{P}_{\nu i}\,(1-\hat{P}_{\nu i})$. Per confrontare le $MS(WT)_i$ di differenti *item* Wright e Masters (1982) propongono una sua trasformazione considerando la

$$t_i^{'} = (MS(WT)_i^{1/3} - 1)(3/sd[MS(WT)_i]) + (sd[MS(WT)_i]/3) \tag{2.71}$$

che ha un valor atteso prossimo allo zero e una deviazione standard prossima ad uno.

L'*item fit mean square* (*INFIT* e *OUTFIT*) è diventato lo strumento sta-

tistico principale per valutare la "bontà" delle domande che compongono un test, così come il suo analogo *person fit mean square* è stato adottato quasi da tutti gli studiosi come strumento per valutare l'adattamento al modello delle risposte dei singoli individui. Nonostante il loro diffuso utilizzo l'*item fit mean square* e il *person fit mean square* sono indicatori poco precisi e poco affidabili, le cui proprietà e le cui distribuzioni, sotto diverse ipotesi di violazione del modello, rimangono ancora ignote.

Innanzi tutto i residui standardizzati non sono lineari da cui deriva che tutte le statistiche di adattamento basate sugli $Z_{\nu i}$ (quindi gli INFIT, gli OUTFIT e tutte le loro trasformazioni) non sono lineari. L'utilità dell'analisi delle componenti principali dei residui proposta da Linacre (1998) per individuare più dimensioni in un test risulta fortemente inficiata dal fatto che gli $Z_{\nu i}$ non rappresentano quantità misurabili su una scala di tipo intervallare ma sono z-score di tipo ordinale.

Come ha poi osservato Richard Smith (1988, 1991), il fatto che le risposte osservate $X_{\nu i}$ siano usate per stimare sia i parametri degli item e degli individui sia i valori attesi $\widehat{P}_{\nu i}$ comporta una sottostima degli $Z_{\nu i}$ e, di conseguenza, una minor possibilità di rilevare i casi di *misfit*.

Studi di simulazione (Smith-Scumacker-Bush, 1998, Karabatsos, 2000), inoltre, hanno evidenziato che, nell'ipotesi in cui i dati si adattino al modello, le distribuzioni degli $Z_{\nu i}$ variano in funzione di numerosi fattori arbitrari; se da una parte il numero di soggetti esaminati e la lunghezza del test influiscono in maniera pressoché irrilevante sulla media dei *mean squares*, dall'altra condizionano molto (soprattutto l'ampiezza del campione) la loro *standard deviation*; in particolare, sotto molti scenari, la deviazione standard degli *OUTFIT* è approssimativamente doppia di quella degli *INFIT*.

Inoltre la distribuzione dei *mean squares* non è simmetrica attorno a 1; valori estremi inferiori all'unità si riscontrano molto meno frequentemente rispetto a quelli superiori a 1, da cui segue che adottare valori critici simmetrici per valutare il *misfit* di un item conduce a errori di I specie differenti nell'una e nell'altra coda della distribuzione. Poichè la distribuzione dei *mean squares* varia a seconda delle condizioni, non è possibile considerare prefissati valori critici costanti (solitamente minori di 0.7 e maggiori di 1.3 oppure minori di

0.8 e maggiori di 1.2) per accettare o meno l'adattamento dei dati al modello. Smith (1998) ha trovato, ad esempio, che a un valore critico pari a 1.2 corrisponde un errore di I tipo effettivo che varia da 0.00001 a 0.10, a seconda della struttura di riferimento considerata. Se si considerano invece le trasformate t e t' con valori critici pari a ± 2, gli errori di I tipo per la statistica non pesata sono approssimativamnete doppi di quelli relativi alla statistica pesata, sebbene la differenza tra le due versioni siano meno accentuate di quelle che si riscontrano tra le due versioni dei *mean squares* e le statistiche t e t' siano meno influenzate dall'ampiezza del campione. Ciò nonostante Karabatsos privilegia l'utilizzo di test di adattamento svincolate dagli *standard residuals* proponendo statistiche basate sul confronto tra risposte osservate e risposte appartenenti alla matrice di Guttman sottostante.

Capita spesso, nella pratica, che il numero di individui per ogni score osservato non sia sufficiente per determinare con soddisfacente affidabilità i valori delle statistiche di *INFIT* e *OUTFIT*. Per ovviare a tale inconveniente Andrich (1982b) ha proposto una statistica, la Stima basata sull'interazione Item-Tratto (*Item-Trait Interaction Test of Fit*), basata sul confronto tra le risposte osservate e i valori attesi di classi di individui. Per determinare tale statistica il campione dei soggetti viene suddiviso in G classi (*Class Intervals*) raggruppando tra loro valori adiacenti delle stime ordinate della loro abilità. Il numero G delle classi viene determinato in base alla numerosità del campione e alla eterogeneità dei punteggi totali, cercando di costruire classi che abbiano dimensioni simili. Nello specifico si considera la quantità:

$$Z_{gi} = \frac{\sum_{\nu \in g} X_{\nu i} - \sum_{\nu \in g} E[X_{\nu i}]}{\sqrt{\sum_{\nu \in g} V[X_{\nu i}]}} \tag{2.72}$$

dove il valore del punteggio totale, $X_{gi} = \sum_{\nu \in g} X_{\nu i}$, è la somma delle risposte corrette di tutti gli individui la cui abilità cade nello specificato intervallo e il valore atteso $E[X_{\nu i}]$ e la varianza $V[X_{\nu i}]$ sono calcolati dapprima per ogni interazione persona-item e poi sommati su tutte le persone con abilità stimata che cade nell'intervallo g. Il valore di Z_{gi} elevato al quadrato costituisce la

componente della statistica

$$\chi_i^2 = \sum_g Z_{gi}^2 = \sum_g [\frac{(\sum_{\nu \in g} X_{\nu i} - \sum_{\nu \in g} E[X_{\nu i}])^2}{\sum_{\nu \in g} V[X_{\nu i}]}] \tag{2.73}$$

che si distribuisce come una chi-quadrato con (G - 1) gdl. In RUMM2020 questa statistica viene utilizzata per misurare le discrepanze tra le percentuali di risposta corretta e le percentuali teoriche stimate dal modello per ogni singolo item e va sotto il nome di *Individual Item Fit*.

Poiché la potenza di rilevare *misfit* usando la statistica del chi-quadrato dipende dall'ampiezza del campione, ogni analisi ha una potenza differente. In molte situazioni con un numero elevato di osservazioni la potenza del test è così elevata che tutti gli item risultano avere un *misfit*, anche quando i valori osservati ed i valori attesi sono prossimi gli uni agli altri. Questo fatto risulta chiaro riscrivendo l'*Individual Item Fit* come segue:

$$\chi_i^2 = \sum_g \left[\frac{\left(\frac{n_g \sum\limits_{\nu \in g} X_{\nu i}}{n_g} - \frac{n_g \sum\limits_{\nu \in g} E[X_{\nu i}]}{n_g} \right)^2}{\frac{n_g \sum\limits_{\nu \in g} V[X_{\nu i}]}{n_g}} \right]$$

$$= \sum_g \frac{(n_g \bar{X}_{gi} - n_g \bar{E}[X_{gi}])^2}{n_g \bar{V}[X_{gi}]}$$

$$= \sum_g [\frac{n_g^2 (\bar{X}_{gi} - \bar{E}[X_{gi}])^2}{n_g \bar{V}_{gi}}]$$

$$= \sum_g [\frac{n_g (\bar{X}_{gi} - \bar{E}[X_{gi}])^2}{\bar{V}[X_{gi}]}]$$

Così ipotizzando che l'ampiezza del campione aumenti mentre i valori medi delle classi rimangono sostanzialmente invariati, il valore del χ_i^2 aumenta in funzione dell'ampiezza del campione n. Studi di simulazione hanno evidenziato che significativi cambi di valore del χ_i^2, all'aumentare della numerosità campionaria, si riscontrano per gli item che non fittano con il modello. In RUMM2020 esiste un'opzione che permette di assegnare una dimensione cam-

pionaria arbitraria per il calcolo del χ_i^2, in modo tale da poter analizzare il comportamento dell'*Individual Item fit* per differenti valori di n.

In modo analogo un test di adattamento globale che confronta le probabilità stimate con le proporzioni osservate può esser costruito mediante la statistica che utilizza i residui standardizzati espressi dalla 2.64

$$\chi^2 = \sum_{r=1}^{k-1} \sum_{i=1}^{k} \frac{(n_{ri} - n_r \hat{P}_{ri})^2}{n_{ri} \hat{P}_{ri}(1 - \hat{P}_{ri})}$$

che si distribuisce approssimativamente come una chi-quadrato con (k-1)(k-2) gdl. Come nel caso dell'analisi del *fit* degli item, con molti item e una moderata ampiezza campionaria, capita abbastanza spesso di avere *total scores* con una bassa frequenza osservata. Anche in questo caso la statistica viene calcolata formando G intervalli di abilità adiacenti e la statistica si distribuisce come una chi-quadrato con (G - 1)(k - 1) gdl.

È possibile verificare l'ipotesi che le stime condizionate dei parametri degli *item* siano indipendenti dall'abilità degli individui con test basati sul Rapporto di Verosimiglianza (Andersen 1973b). I parametri degli item sono stimati prima sull'intero campione e successivamente su ogni gruppo che ha totalizzato un *total score* pari a r. Ponendo $L(\hat{\delta}_i)$ la verosimiglianza condizionata con il parametro stimato sull'intero campione e con $L(\hat{\delta}_i^{(r)})$ la stessa verosimiglianza con le stime valutate solo sul gruppo di *total score* uguali a r, segue che la statistica

$$-2ln[\frac{L(\hat{\delta}_i)}{\prod_r L(\hat{\delta}_i^{(r)})}] \tag{2.74}$$

si distribuisce come una v.c. chi-quadrato con (L-1)(L-2) gdl. Anche in questo caso se la numerosità nei diversi gruppi di *score* è piccola, si usa raggruppare tra loro classi di punteggio adiacenti in G gruppi ottenendo una variabile che si distribuisce come una chi-quadrato con (G-1)(L-1) gdl.

2.4.2 Invarianza delle stime dei parametri fra sottogruppi: l'analisi del DIF

L'invarianza delle stime degli item può essere verificata, oltre che tra gli individui appartenenti a intervalli distinti di abilità, anche tra sottogruppi del campione definiti secondo una classificazione di una caratteristica dei soggetti presente all'interno della struttura di riferimento (sesso, età, condizione socio-economica, area geografica di residenza, ecc.). Può infatti accadere che ogni sottogruppo costituisca una struttura di riferimento a se stante e che solo all'interno di ciascuna struttura di riferimento sia possibile costruire una misura che goda dell'*oggettività specifica* mentre sia impossibile costruirla sull'intera popolazione o sull'intero campione.

Succede anche frequentemente, nelle applicazioni pratiche, che solamente alcuni *item* si comportino in maniera sistematicamente differente (pur rispettando i vincoli del RM) in ciascun sottogruppo e che manifestino invece un malfunzionamento (violando i requisiti richiesti dal modello) se valutati sull'intera popolazione o sull'intero campione. È il caso del *Differential Item Functioning*, in base al quale uno stesso item può presentare *differenze di intensità*, variando la sua difficoltà all'interno della scala tra un sottogruppo e l'altro, o *differenze di genere*, variando il suo comportamento tra un sottogruppo e l'altro.

La presenza del *DIF* in uno o più item costituisce una violazione della condizione dell'*oggettività specifica* complessiva, in quanto la probabilità di rispondere correttamente a questi item, per lo stesso livello di abilità, varia tra i soggetti dei diversi sottogruppi; la soluzione, il più delle volte, consiste nel dividere l'item (o gli item) *biased* in tanti nuovi item quanti sono i sottogruppi di persone in cui si manifesta un comportamento differente e assegnare ognuno di questi nuovi item al suo sottogruppo corrispondente (ponendo un valore mancante a tutti i sottogruppi a cui non viene assegnato).

Una statistica per valutare la differenza di intensità di un item tra due gruppi di soggetti è costituita dalla quantità:

$$Z_i = \frac{\hat{\delta}_{iA} - \hat{\delta}_{iB}}{[\hat{\sigma}^2_{iA} + \hat{\sigma}^2_{iB}]^{\frac{1}{2}}}. \tag{2.75}$$

Se la numerosità è sufficientemente elevata la Z_i si distribuisce secondo una Normale standardizzata; allora la statistica $\chi^2 = \sum_i Z_i^2$ si distribuisce approssimativamente secondo una v.c. chi-quadrato con (*k-1*) gdl e può essere utilizzata per testare significative differenze di scala tra sottogruppi (Van den Wollenberg, 1982). Generalizzazioni di questo test sono state sviluppate da Andrich (Andrich and Kline, 1981) per analizzare il *DIF* di uno o più item, considerando contemporaneamente classificazioni degli individui rappresentabili con tabelle a una o a due vie.

Anche la statistica 2.74 può essere adattata per testare la differenza delle stime di item tra sottogruppi di persone, considerando anziché la verosimiglianza calcolata sul sottogruppo con *total score* uguale a r, quella sul sottogruppo con un certo valore di una determinata caratteristica.

Le tecniche per analizzare il *DIF* basate sull'analisi dei residui (Wright e Stone, 1979; Andrich 1982a, Andrich e Hagquist, 2004) sono molto diffuse, sia per la loro semplicità concettuale sia per l'attendibilità dei risultati. Per ogni item viene stimato un set di parametri e successivamente si studiano i residui ottenuti nei diversi sottogruppi di persone, aggregate rispetto a una o più caratteristiche. In RUMM2020 il confronto tra residui viene valutato con l'Analisi della Varianza (ANOVA) che scompone la varianza totale dei residui standardizzati

$$Z_{n_{cg}i} = \frac{X_{n_{cg}i} - E[X_{n_{cg}i}]}{\sqrt{Var(X_{n_{cg}i})}} \tag{2.76}$$

per testare la significatività:

i dell'effetto principale del fattore (*main group effect*);

ii dell'effetto della *CI* (*main class effect*);

iii dell'effetto prodotto dall'interazione tra i due (*group-by-class interval effect*).

Gli $Z_{n_{cg}i}$, analoghi ai residui espressi dalla 2.63, hanno un indice c che rappresenta la classe intervallo (*CI*) d'appartenenza e un indice g che rappresenta il livello del fattore possibile causa del *DIF*,

Nell'analisi del *DIF* quello che più interessa studiare è se la media dei residui risulta significativamente diversa (i) e se l'effetto discriminante dell'item

è significativamente diverso (iii) fra i gruppi. In questo ambito l'analisi a livello di classe (ii) risulta di secondaria importanza in quanto si riferisce alle medie delle abilità nelle classi in cui è stato suddiviso il campione osservato e non fornisce alcuna informazione aggiuntiva rispetto all'analisi del *fit* degli item, già trattata in precedenza con l'*Item-Trait Fit Statistics* basata sui residui espressi dalla 2.72.

Tale metodo di verifica ha trovato molta adesione tra gli studiosi del RM; sebbene i residui standardizzati non siano lineari e solo approssimativamente distribuiti come una Normale, l'affidabilità della distribuzione F dell'ANOVA nel testare l'assenza di *DIF* e il buon funzionamento generale dell'item è molto alta. Ogni valutazione di tipo statistico va comunque sempre accompagnata dall'osservazione delle *ICC* e da considerazioni metodologiche riguardanti la costruzione degli item.

2.4.3 Generalizzazioni del Test del chi-quadrato di Pearson e altri test di adattamento

Si discutono ora alcune generalizzazioni del test del chi-quadrato di Pearson (Fischer e Molenaar; 1995) considerando un modello multinomiale con M risultati possibili, mutualmente esclusivi, con probabilità $\pi_1(\phi)$, $\pi_2(\phi)$, ..., $\pi_M(\phi)$, dove ϕ denota il vettore q-dimensionale dei parametri del modello. Siano p_1, p_2, ..., p_M le proporzioni osservate e $\hat{\pi}_j$ le $\pi_j(\phi)$ valutate con le stime di massima verosimiglianza di ϕ. Sotto condizioni di regolarità molto generali, la statistica

$$\chi^2 = n \sum_{j=1}^{M} \frac{(p_j - \hat{\pi}_j)^2}{\hat{\pi}_j} \tag{2.77}$$

si distribuisce asintoticamente, per $n \to +\infty$, come una v.c. chi-quadrato con *M-q-1* gdl. Nei modelli *IRT* la sommatoria comprende tutti i 2^k possibili *pattern* di risposta ma ciò comporta a) che il numero di frequenze coinvolto è troppo grande perché la statistica possa essere informativa contro le violazioni del modello e b) che le frequenze attese sono molto piccole e inficiano l'ipotesi sulla distribuzione asintotica della statistica.

Glas e Verhelst (1989) hanno introdotto una classe di test statistici asintoticamente distribuiti come chi-quadrati che superano questo problema. Ponendo $\boldsymbol{p} = (p_1, \ldots, p_M)'$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)'$, $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_M)'$ e $\mathbf{b} = \mathbf{n^{1/2}}(\mathbf{p} - \hat{\boldsymbol{\pi}})$ la 2.77 si può riscrivere come:

$$\chi^2 = \mathbf{b}'\hat{\mathbf{D}}_{\boldsymbol{\pi}}^{\mathbf{-1}}\mathbf{b}$$

dove $\boldsymbol{D_\pi}$ è la matrice diagonale M x M degli elementi di $\boldsymbol{\pi}$. Una classe molto ampia di test di Pearson si basa su un vettore di G combinazioni lineari $\mathbf{d} = \mathbf{U}'\mathbf{b}$ dove la matrice dei contrasti $\mathbf{U}$ di dimensioni M x G è scelta in maniera tale che $G << M$ e le combinazioni lineari possano evidenziare specifiche violazioni del modello, utilizzando probabilità attese sufficientemente ampie da poter applicare la teoria asintotica. Si prenda in considerazione la statistica:

$$Q = Q(\mathbf{U}) = \mathbf{b}'\mathbf{U}(\mathbf{U}'\hat{\mathbf{D}}_{\pi}^{\mathbf{-1}}\mathbf{U})^{-}\mathbf{U}'\mathbf{b} = \mathbf{d}'\mathbf{W}^{-}\mathbf{d} \tag{2.78}$$

dove $(\mathbf{U}'\hat{\mathbf{D}}_\pi\mathbf{U})^-$ e $\mathbf{W}^-$ sono, rispettivamente, le inverse generalizzate di $(\mathbf{U}'\hat{\mathbf{D}}_\pi\mathbf{U})$ e $\mathbf{W}$ $\mathbf{d}$ è il vettore delle *distanze* e $\mathbf{W}$ la matrice dei pesi. Glas e Verhelst (1989) hanno trovato condizioni sufficienti affinché la 2.78 sia distribuita asintoticamente come una v.c. chi-quadrato con gdl = rango($\mathbf{U}'\mathbf{D}_\pi\mathbf{U}$) - q - 1. Una condizione è la seguente.
Sia $\mathbf{A}$ una matrice M x q $\{\{a_{mj}\}\}$ con

$$a_{mj} = \pi_m^{-1/2}\frac{\partial \pi_m}{\partial \phi_j} \tag{2.79}$$

tale che $\mathbf{A} = \mathbf{D}_\pi^{\mathbf{-1/2}}\partial\pi_{\mathbf{m}}/\partial\phi_{\mathbf{j}}$.
Per ogni arbitraria $\mathbf{P}$, sia $\wp(\mathbf{P})$ l'insieme delle combinazioni lineari delle colonne di $\mathbf{P}$. Allora Q($\mathbf{U}$) si distribuisce asintoticamente come una v.c. chi-quadrato se sono soddisfatte le due seguenti condizioni:

1. Le colonne di $\mathbf{A}$, definita dalla 2.79, appartengono a $\wp(\mathbf{D}_\pi^{\mathbf{1/2}}\mathbf{U})$;
2. Esiste un vettore di costanti $\mathbf{c}$ tale che $\mathbf{Uc} = \mathbf{1}$, con $\mathbf{1}$ vettore M dimensionale di 1.

Come si costruisce $\mathbf{U}$? Si può prendere in considerazione una certa $\mathbf{U}$ di contrasti e verificare che essa verifichi la 1) e la 2) oppure si sviluppa una procedura

per costruirla. Nel caso di modelli appartenenti alla famiglia esponenziale tale procedura è relativamente semplice.

Il test R_1 e il test Q_1 mirano a verificare la monotonicità e il parallelismo delle *ICC* del RM. Gli score $1, \ldots, k-1$ degli individui sono divisi in G sottogruppi a seconda del punteggio (oppure in base all'abilità latente stimata, o secondo una variabile di controllo esterna) e ogni individuo è assegnato a un sottogruppo in base allo *score* totalizzato.

Sia M_{1gi}, con realizzazione m_{1gi}, il numero di persone appartenenti alla classe g che hanno risposto positivamente all'item I_i e sia $E(M_{1gi}|\hat{\theta}, \hat{\delta})$ il suo valore atteso condizionato, cioè il suo valore atteso, data la distribuzione di frequenza degli *score* dei rispondenti e le stime di massima verosimiglianza condizionata dei parametri degli item. I due test si basano sulle differenze:

$$d^*_{1gi} = m_{1gi} - E(M_{1gi}|\hat{\theta}, \quad \hat{\delta}) \tag{2.80}$$

divise per la loro *standard deviation* stimata al fine di ottenere le variabili binomiali standardizzate z_{1gi}. Calcolando il quadrato degli z_{1gi} e sommando tali valori su gruppi opportuni, si definisce un indice globale dell'adattamento dei dati al modello la cui distribuzione può essere determinata tenendo in considerazione la dipendenza tra le d^*_{1gi}.
Una statistica è espressa dalla:

$$R_{1c} = \sum_g \mathbf{d}'_{\mathbf{1g}} \mathbf{W}^-_{\mathbf{1g}} \mathbf{d}_{\mathbf{1g}} \tag{2.81}$$

con $\mathbf{d_{1g}}$ vettore di elementi $d_{1gi} = d^*_{1gi}/\sqrt{n}$, per tutti gli item del test, e $\mathbf{W_{1g}}$ è la matrice dei pesi $\mathbf{U}'\hat{\mathbf{D}}_\pi\mathbf{U}$. Tale matrice è uguale alla matrice di covarianza delle $\mathbf{d_{1g}}$ (Glas, 1988a, 1989) e la statistica R_{1c} ha una distribuzione asintotica chi-quadrato con (G - 1)(k - 1) gdl.
Un'altra statistica basata sulle d^*_{1gi} è il Q_1 di Van de Wollenberg (1982), definita dalla

$$Q_1 = \frac{k-1}{k} \sum_{i=1}^{k} \sum_{g=1}^{G} z^2_{1gi} \tag{2.82}$$

Anche Q_1 si distribuisce approssimativamente come una chi-quadrato con (G -

1)(k - 1) gdl. Q_1 può essere interpretato come un'approssimazione di R_{1c} dove la matrice dei pesi è sostituita da una matrice diagonale con elementi uguali agli elementi della diagonale della matrice di covarianze completa.

Altri test si focalizzano esplicitamente sul comportamento degli item. In particolare si possono utilizzare le z_{1gi} per valutare variazioni dell'*item discrimination*. Molenaar (1983) ha elaborato una statistica il cui segno indica se un item discrimina troppo o troppo poco. Siano c_1 e c_2 due valori che suddividono il range degli *score* in tre regioni, una bassa, una media ed una alta. Poiché le z_{1gi} sono binomiali standardizzate la statistica:

$$U_i = \frac{\sum_{g=1}^{c_1} z_{1gi} - \sum_{g=c_2}^{k-1} z_{1gi}}{(c_1 + k - c_2)^{1/2}} \tag{2.83}$$

si distribuisce approssimativamente come una normale standardizzata. Generalmente le due soglie c_1 e c_2 sono scelte in modo che entrambe le sommatorie includano il 25% degli individui del campione. Se la funzione delle risposte osservate è più inclinata rispetto a quella teorica del modello, la U_i è negativa; se, invece, è meno inclinata, la U_i è positiva. Si può ricondurre questa statistica ai test generalizzati di Pearson elevando al quadrato le d_{1gi} che sottostanno al calcolo degli z_{1gi} e riscalandoli con una matrice di varianza-covarianza per ottenere una statistica che si distribuisce asintoticamente come una chi-quadrato con 1 gdl.

Una statistica analoga, la S_i, è stata sviluppata da Verhelst ed Eggen (1989). Anch'essa, come la statistica R_{1c}, si basa sulle d_{1gi} ed è efficace nel rilevare le differenze di discriminazione tra gli item. Il range degli *score* è suddiviso in $g = 1, ..., G$ classi equivalenti e per ogni livello di *score* g, si calcola la differenza tra il numero di risposte corrette osservate e il numero di risposte corrette attese per un item. Allora la statistica:

$$S_i = \mathbf{d_i'} \mathbf{W_i^-} \mathbf{d_i} \tag{2.84}$$

ha una distribuzione asintotica chi-quadrato con G - 1 gdl.

Altre statistiche definite per testare la presenza di violazioni dell'unidimensionalità del costrutto mirano a verificare l'associazione tra gli item. Infatti,

in caso di più dimensioni, la posizione dell'individuo rispetto al tratto latente non è sufficientemente descritta dal parametro unidimensionale che definisce la variabile oggetto di studio e, di conseguenza, l'associazione tra le risposte agli item, dato tale parametro, non scompare come dovrebbe, valendo le ipotesi del modello.

Sia M_{2ij} il numero di individui che ottengono uno *score* maggiore di uno, avendo risposto correttamente sia a I_i che a I_j e sia $E(M_{2ij}|\hat{\theta}, \hat{\delta})$ il valore atteso di massima verosimiglianza condizionata e m_{2ij} una sua realizzazione. La differenza

$$d^*_{2ij} = m_{2ij} - E(M_{2ij}|\hat{\theta}, \quad \hat{\delta}) \tag{2.85}$$

può essere utilizzata per individuare violazioni dell'unidimensionalità. Si consideri la forma quadratica:

$$\mathbf{d_2^{'} W_2^{-1} d_2} \tag{2.86}$$

dove $\mathbf{d_2}$ è il vettore delle differenze $d^*_{2ij}/\sqrt{(n)}$ di tutte le coppie di item e $\mathbf{W_2^{-1}}$ l'inversa della sua matrice di covarianza stimata.
Sia $\mathbf{d_1}$ un vettore che ha come elementi, per tutti gli item, le differenze tra il numero di persone che rispondono correttamente all'item e ottengono uno *score* uguale a uno e il suo valore atteso condizionato tramite la CML, entrambi divisi per la radice quadrata di n. Sia $\mathbf{W_1^{-1}}$ la sua matrice di covarianza stimata. Allora

$$R_{2c} = \mathbf{d_2{}' W_2^{-1} d_2} + \mathbf{d_1{}' W_1^{-1} d_1} \tag{2.87}$$

ha una distribuzione asintotica chi-quadrato con $k(k-1)/2$ gdl. Anche questa statistica appartiene alla classe dei test generalizzati di Pearson (Glas, 1989).

Nella statistica Q_2 di Van den Wollenberg (1982) il campione di persone a cui è sottoposto il test è suddiviso in G sottogruppi in base allo *score* ottenuto o a variabili esterne. Suddividendo ulteriormente le differenze d^*_{2ij} si definisce M_{2gij}, il numero di persone nel sottogruppo g che risponde correttamente sia a I_i che a I_j. La differenza tra la sua realizzazione e il suo valore atteso stimato con la *CML* è, analogamente alla 2.85:

$$d_{2gij} = m_{2gij} - E(M_{2gij}|\hat{\theta}, \quad \hat{\delta}). \tag{2.88}$$

Dividendo queste differenze per la loro *standard deviation* stimata si ottengono le variabili standardizzate z_{2gij} per la formula:

$$Q_2 = \frac{k-3}{k-1} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sum_{g=1}^{G} z_{2gij}^2. \tag{2.89}$$

Se i parametri degli item sono stimati in ogni sottogruppo, la distribuzione di Q_2 approssima quella di una v.c. chi-quadrato con $G\ k\ (k\text{-}3)/2$ gdl.

Capitolo 3

L'analisi delle prove dell'INVALSI

3.1 La pulizia dei database

L'utilizzabilità delle informazioni contenute nei database con i risultati dei test relativi alle classi II e IV elementare degli anni scolastici 2004/2005 e 2005/2006 del Sistema Nazionale di Valutazione (SNV) è stata inficiata dal fatto che i dati erano considerati scarsamente affidabili per questioni inerenti la *reliability*, il *teacher cheating*, la qualità dei questionari, ecc. La prima verifica empirica effettuata sui dati è stata l'analisi delle distribuzioni degli *score* degli studenti che ha fatto sorgere il forte sospetto sull'esistenza di problemi legati al *teacher cheating* (l'influenza dell'intervento dell'insegnante sui risultati delle prove somministrate); infatti si è riscontrata una differenza sostanziale tra le distribuzioni dei punteggi nelle diverse aree geografiche (province, regioni e macro aree); in alcune zone le distribuzioni hanno un andamento pressoché normale, in altre, invece, presentano una forte asimmetria a sinistra come si può constatare osservando i grafici 3.1 e 3.2 ottenuti dai dati originali standardizzati suddivisi nelle cinque macro aree italiane (Nord Ovest: Piemonte, Valle d'Aosta, Liguria e Lombardia; Nord Est: Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna; Centro: Toscana, Umbria, Marche, Lazio; Sud: Abruzzo, Molise, Campania e Puglia; Sud e Isole: Basilicata, Calabria,

Sicilia e Sardegna).

Per poter sfruttare il contenuto informativo dei test relativi alle classi IV degli anni scolastici 2004/2005 e 2005/2006 è stato individuato e implementato un protocollo per la pulizia dei dati che verrà descritto nel presente capitolo. Tale protocollo si basa:

- sull'analisi del pretest (nel caso questo sia stato somministrato);
- sugli indicatori di *cheating* (Levitt, 2002);
- sull'analisi degli item dei questionari;
- sull'analisi del *misfit* delle osservazioni rispetto al Modello di Rasch.

Nelle intenzioni l'analisi del pretest voleva raffrontare il comportamento degli item del test SNV 2004/2005 confrontandolo con quello riscontrato nel corrispondente pretest e il comportamento degli item del test SNV 2005/2006 con quello avuto nei pretest. Purtroppo i dati del pretest della prova SNV 2004/2005 sono risultati irreperibili mentre, poichè per la somministrazione delle prove avvenuta nell'anno scolastico 2005/2006 non è stata effettuata nessuna operazione di pretesting, per verificare l'eventuale consonanza delle domande già somministrate precedentemente i risultati degli item delle prove oggetto di studio sono stati confrontati con quelli ottenuti nel prove di valutazione dell'anno precedente. Sia per gli item identici che per quelli simili è stato verificato se il livello di difficoltà nel pretest è stato conservato nelle prove del SNV vere e proprie.

Mediante la costruzione e l'utilizzo di indicatori di *cheating* si sono analizzati i *pattern* di risposta per individuare gruppi omogenei di studenti (intere classi o intere scuole) in cui fosse evidente l'intervento dell'insegnante nello svolgimento della prova.

Infine, con l'analisi sulla bontà degli item e l'analisi del *misfit* delle risposte si sono voluti individuare gli item e gli scolari che presentavano un cattivo adattamento al RM violando, in maniera più o meno significativa, gli assunti alla base del modello.

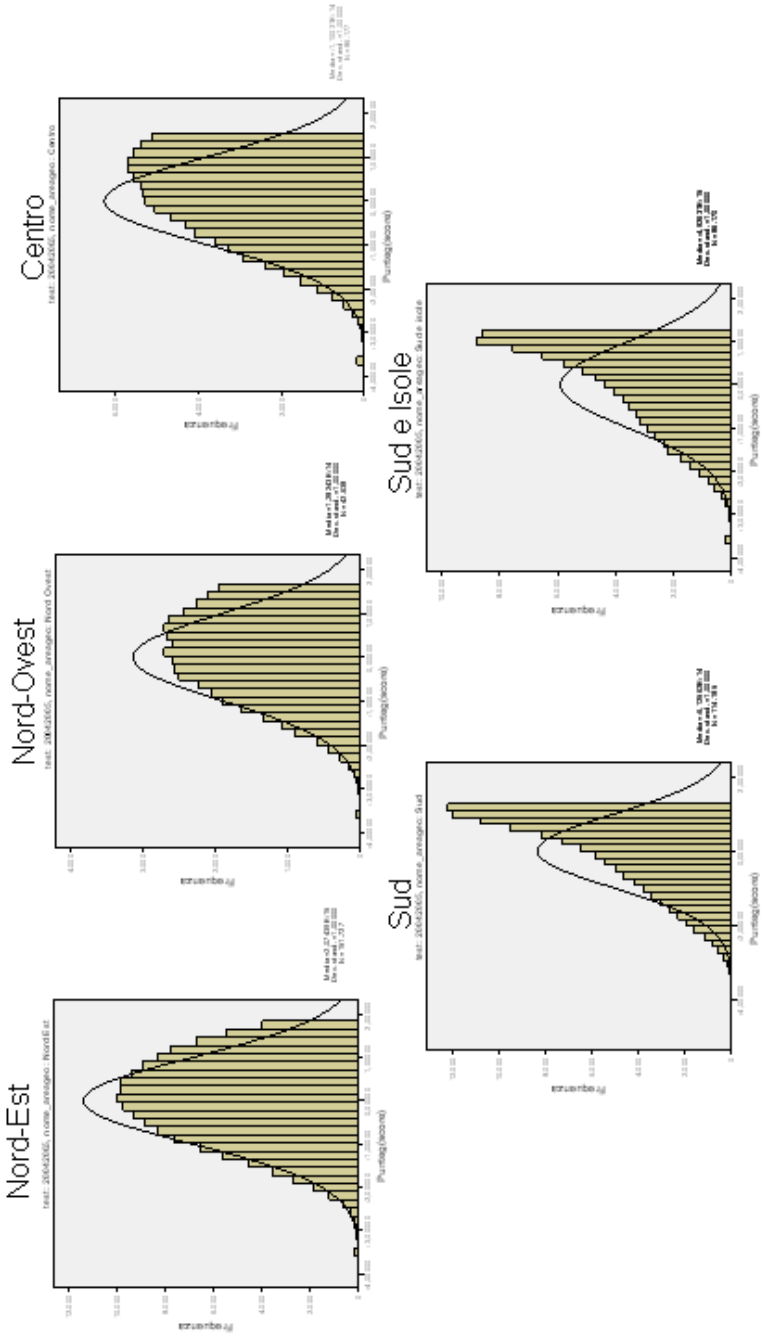


Figura 3.1: Distribuzioni dei total score standardizzati nel 2004/2005

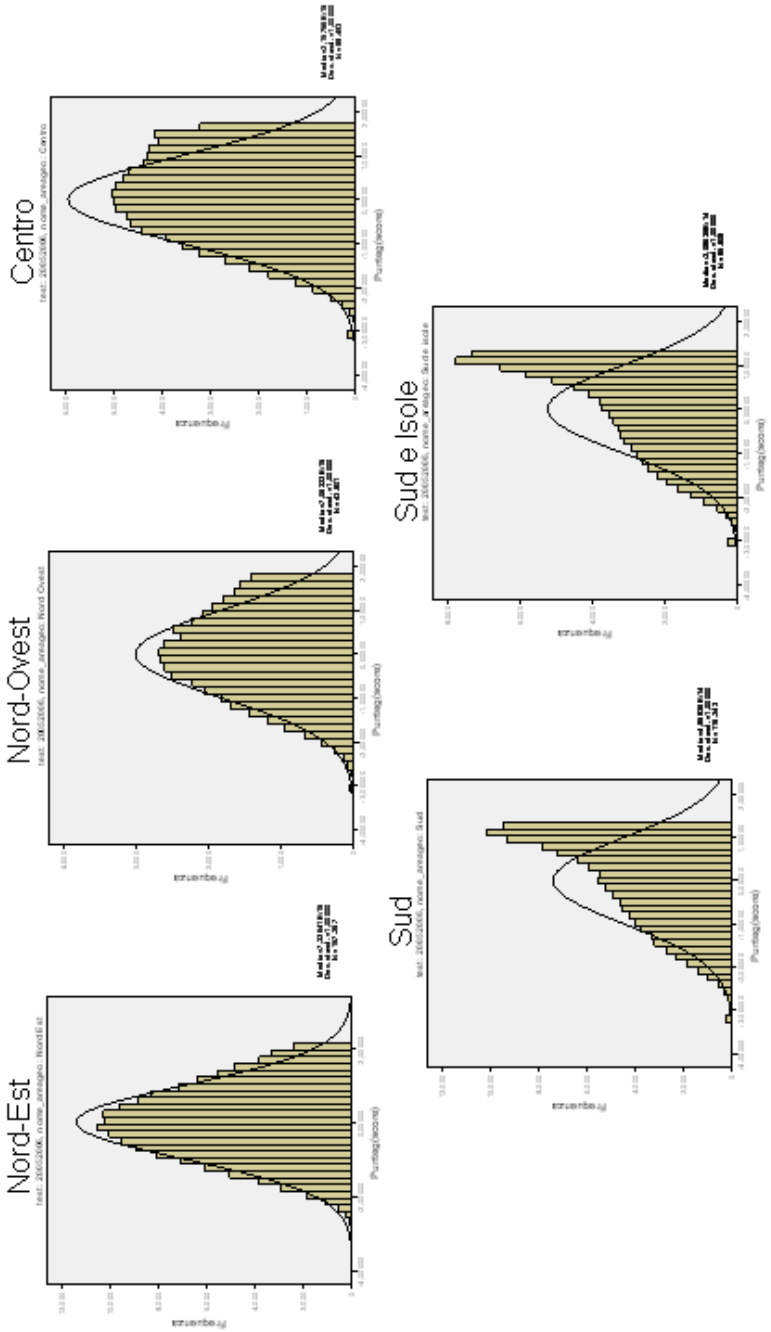


Figura 3.2: Distribuzioni dei total score standardizzati nel 2005/2006

Per verificare l'adeguatezza e l'efficacia della procedura di pulizia dei dati è stata condotta un'analisi sulla correlazione dei risultati del test del SNV 2005/2006 (prima e dopo la pulizia) con quelli delle prove internazionali TIMSS 2003 e 2007. In effetti la pulizia dei dati ha reso i risultati delle prove del SNV più congruenti con quelli emersi dalle prove TIMSS; ciò non di meno alcuni difetti intrinsechi nel sistema di costruzione e nella modalità di erogazione delle prove (derivanti principalmente da una imprecisa calibratura dei questionari rispetto al livello di abilità da misurare e dal comportamento poco "professionale" di alcuni insegnanti preposti al controllo del corretto svolgimento delle prove) non hanno permesso di ottenere dei risultati apprezzabili.

La fase di pulizia dei dati ha avuto lo scopo di identificare un set di item, in accordo con gli assunti del modello di Rasch, con cui misurare le abilità matematiche negli anni scolastici 2004/2005 e 2005/2006 e da cui poter attingere alcune domande per costruire, con l'aggiunta di altri quesiti presi da altre prove di valutazione (test SNV di altri livelli scolastici e prove TIMSS), il test di aggancio e un set di scolari da cui estrarre un campione per stimare le scale di misurazione dell'apprendimento della matematica in IV elementare.

3.1.1 L'analisi del pretest

La disamina dei questionari per verificare la qualità degli item è cominciata dal confronto delle domande delle prove del SNV con le domande del pretest. Come anzidetto, l'impossibilità di recuperare i risultati del pretest della prova di IV elementare dell'anno scolastico 2004/2005 ha impedito di poter controllare la coerenza del comportamento degli item tra un periodo e l'altro. Invece, poiché sulla prova avvenuta nell'anno scolastico 2005/2006 non è stata effettuata alcuna operazione di pretesting degli item, si sono confrontate le domande somministrate con quelle testate nell'anno precedente in altre prove dell'INVALSI (il pretest nel 2004/2005 è stato effettuato all'inizio dell'anno scolastico nelle classi III e V elementare, mentre la somministrazione è avvenuta all'inizio dell'anno scolastico successivo nelle classi II e IV elementare) verificandone la congruenza in termini di livelli di difficoltà.

A seguito dell'analisi degli item somministrati con la prova del 2005/2006 si è osservato quanto segue:

- era presente 1 item che era già stato somministrato in modo identico nell'anno precedente (item 18);
- era presente 1 item a cui è stato modificato il testo della domanda mantenendo uguali i distrattori (item 16, fig. 3.3);
- erano presenti 3 item a cui sono stati modificati sia il testo che i distrattori (item 5, item 17 e item 21, fig. 3.4 e 3.5);
- non è stato trovato nessun item con testo identico e distrattori modificati;
- i rimanenti 23 item non erano stati testati.

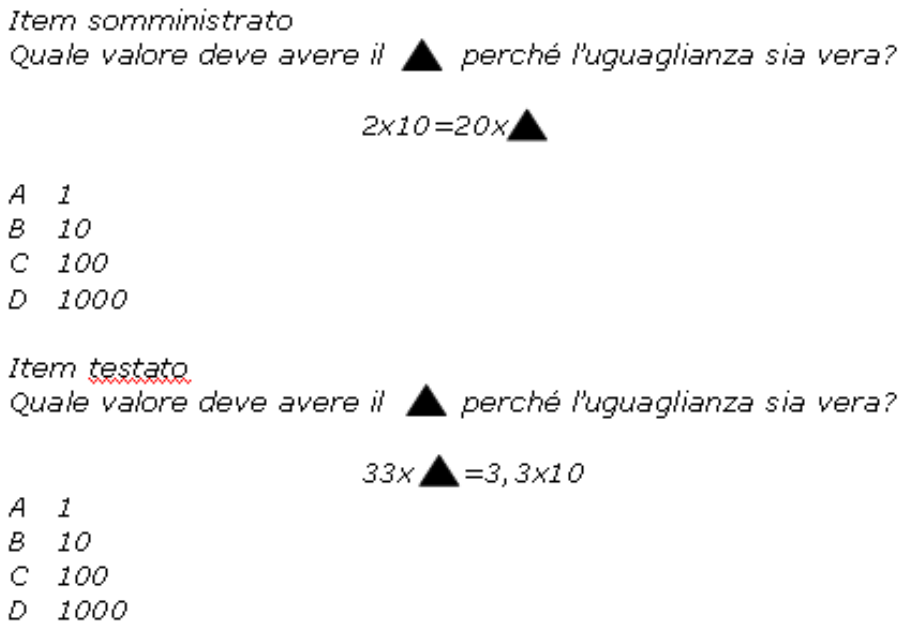
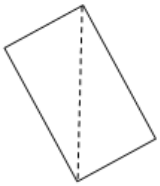


Figura 3.3: Item16 dell'SNV 2005/2006

Item somministrato
Nella figura seguente:

Come si chiama il segmento DH?
A. Altezza
B. Diagonale
C. Base
D. Mediana

Item testato
Come si chiama il segmento BD?



A. Altezza
B. Diagonale
C. Mediana
D. Lato

Figura 3.4: Item5 dell'SNV 2005/2006

3.1.2 Definizione e costruzione degli indicatori di cheating

Per la definizione e la costruzione degli indicatori di *cheating* si è preso come riferimento lo studio condotto da Levitt (2002) su un campione di scuole elementari pubbliche di Chicago mediante il quale è stata sviluppata una procedura per individuare *score* e *pattern* di risposta inattesi. Nel presente lavoro sono stati creati 5 indicatori, quattro (*Ind1-Ind4*) per identificare i casi di pattern di risposta in cui sussiste un forte sospetto di un intervento da parte degli insegnanti e uno (*Ind5*) per identificare i *pattern* di risposta meno plausibili. Tutti e cinque gli indicatori sono dicotomici e assumono valore 1 nel caso in cui le stringhe di risposta risultino positive a uno dei criteri di *cheating* illustrati in seguito e 0 altrimenti; i primi quattro si basano sull'analisi dei valori delle risposte originarie e solo il quinto criterio di individuazione di risposte anomale sfrutta l'analisi di Rasch. Ciò non di meno tutti e cinque i criteri cercano di identificare *pattern* di risposte che violano uno o più degli assunti

Item somministrato
A quale numero corrispondono 3 decine e 12 unità?

A 312
B 42
C 32
D 15

Item testato
A quale numero corrispondono 4 decine, 22 unità e 11 decimi?

A 631
B 422,11
C 63,1
D 42,24

Figura 3.5: Item21 dell'SNV 2005/2006

del RM, quali l'indipendenza tra i soggetti, l'indipendenza di risposta, il *bias* degli item e il *bias* della scala. Infatti tutti e cinque i criteri confrontano tra loro, all'interno della stessa classe, singole stringhe di risposte per individuare possibili analogie o dipendenze tra un *pattern* e l'altro o valutano le discrepanze più marcate tra le singole stringhe osservate e quelle attese per individuare i *pattern* che meno si adattano al modello.

Costruzione del primo Indicatore (*Ind1*)

Con il primo indicatore si individuano le classi in cui tutti gli studenti hanno risposto in modo identico a tutti gli item. Si considera la stringa delle risposte binarie $(x_1, x_2, x_3, \dots)$ di ciascuno studente e si effettua quindi la conversione in numero decimale secondo il metodo noto:

$$x_1 * 2^0 + x_2 * 2^1 + x_3 * 2^2 + \dots$$

Il risultato è un numero univoco mediante il quale è possibile individuare nel database le stringhe identiche indipendentemente dalla correttezza o meno delle risposte. Si calcola la deviazione standard di questa variabile all'interno

di ciascuna classe e quindi si pone *Ind1* uguale a 1 se la deviazione standard è uguale a 0, cioè se nella classe sono state date le stesse identiche risposte a tutti gli item, 0 nel caso in cui la deviazione standard è diversa da 0.

Costruzione del secondo Indicatore (*Ind2*)

Con il secondo indicatore si individuano le classi dove tutti gli studenti hanno risposto correttamente a un certo numero di domande. Si crea un aggregato a livello di classe che controlla il numero di item a cui tutti gli studenti hanno risposto in maniera corretta. A tale scopo si creano le variabili *d1Sum* ... *d28Sum* che sommano, ciascuna, le risposte date all'item corrispondente e che, pertanto, assumono valori che variano da 0 alla numerosità della classe. Si crea quindi un altro gruppo di variabili quanti sono gli item (*d1* ... *d28*) e si attribuisce a ciascuna di esse valore 1 se il valore della variabile corrispondente *d1Sum* ... *d28Sum* è pari alla numerosità della classe; ciò allo scopo di identificare gli item a cui tutti gli studenti della classe hanno risposto in modo corretto. Si calcola la variabile *tot* (somma delle variabili *d1*, ..., *d28*) che indica, all'interno di ogni classe, a quanti item tutti gli studenti hanno risposto in modo corretto; tale variabile può assumere valori compresi tra 0 a 28 (che è il numero complessivo di item sia del test SNV 2004/2005 che nel test SNV 2005/2006). Infine, l'indicatore *Ind2* si pone uguale a 1 se la numerosità della classe è maggiore di 4 e la variabile *tot* è maggiore di 14, cioè se gli studenti di una stessa classe hanno risposto tutti in modo corretto a più di 14 domande, 0 altrimenti.

Costruzione del terzo Indicatore (*Ind3*)

Con il terzo indicatore si individuano le classi in cui l'ordinamento generale delle difficoltà degli item viene sovvertito basandosi sul fatto che la maggior parte degli scolari ha risposto in modo corretto più alle domande difficili che a quelle facili. Dopo aver ordinato gli item secondo la loro difficoltà crescente, per ogni studente si determina nella variabile *primi5* il punteggio ottenuto nei 5 item più facili e nella variabile *ultimi5* il punteggio ottenuto nei 5 item più difficili. Si crea poi la variabile *ind3stu* ponendola uguale ad 1 se *ultimi*5 >

*primi*5 per lo studente in oggetto; si aggrega tale variabile sommandola per classe ottenendo la *ind3stusum* e, infine, si crea l'indicatore *Ind3* che assume valore 1 se *ind3stusum* > Numerosità classe/2, 0 altrimenti.

Costruzione del quarto Indicatore (*Ind4*)

Con il quarto indicatore, costruito a livello di scuola, si individuano le scuole in cui la maggior parte degli scolari sono stati classificati come casi di *cheating* secondo uno dei tre criteri precedenti. Per ogni studente viene creata una variabile che assume valore 1 se la somma *Ind1* + *Ind2* + *Ind3* è diversa da 0 (quindi se quello studente è rientrato in almeno una tipologia di *cheating* definita in precedenza), 0 altrimenti. Si crea una variabile di appoggio che divide il valore 0 o 1 della variabile appena determinata per la numerosità della classe a cui quello studente appartiene e, quindi, si aggrega tale variabile sommandola a livello di scuola. Si pone, infine, *Ind4* uguale a 1 se la somma della variabile appena costruita è maggiore del 50% delle classi di quella scuola (nel qual caso viene eliminata dal database l'intera scuola), 0 altrimenti.

Costruzione del quinto Indicatore (*Ind5*)

Con il quinto indicatore si va a osservare, all'interno di ogni classe, la percentuale di studenti che presentano una condizione di *misfit* secondo gli assunti del modello di Rasch (come criterio generale si sono indicati come soggetti con un cattivo adattamento tutti gli studenti con un $infit < 0.75$ o $infit > 1,25$). Si è posto *Ind5* uguale a 1 se più del 50% degli studenti di una classe ha un *misfit* (e la classe viene rimossa dal database), 0 altrimenti.

3.1.3 La procedura di pulizia dei database

1. Nel database originale contenente tutti i dati è stato determinato, all'interno di ogni classe, il numero di stringhe identiche indipendentemente dal numero di risposte corrette, sono stati determinati *Ind1* e la numerosità della classe e successivamente si sono contrassegnate nei database tutte le classi con numerosità superiore a 4 in cui *Ind1* = 1.

2. Si è considerata, poi, la presenza di classi nelle quali tutti gli studenti hanno risposto correttamente allo stesso numero di domande e si è posto *Ind2* uguale a 1 se gli studenti hanno risposto correttamente a più di 14 domande identiche; nel database sono state contrassegnate le classi in cui *Ind2* = 1 e la numerosità della classe era maggiore di 4.

3.
   - A questo punto, dalla popolazione depurata dai casi in cui *Ind1 = 1* e *Ind2 = 1*, sono stati estratti due campioni ciascuno di numerosità pari a 50.000 studenti: un campione ufficiale *Camp1*, stratificato per Provincia, con allocazione proporzionale negli strati, e un campione di controllo *Camp2* non stratificato;
   - i dati di ciascun campione sono stati analizzati con il modello di Rasch, utilizzando i due software RUMM2020 e Winsteps, allo scopo di determinare la difficoltà degli item e individuare quelli che presentavano un cattivo adattamento (*misfit*);
   - si è applicata alla popolazione totale l'analisi di Rasch effettuata con Winsteps per determinare gli stessi parametri a livello globale ed effettuare un confronto con i campioni.

4. Stimate le difficoltà degli item si sono identificati i casi di inversione di risposte tramite il terzo indicatore (*Ind3*) e si sono considerate classi "anomale" tutte quelle in cui il punteggio delle prime 5 domande più facili è stato strettamente inferiore al punteggio delle ultime 5 domande più difficili, anche in questo caso considerando sempre classi con numerosità superiore a 4 studenti.

5. Successivamente si sono identificate, tramite il quarto indicatore *Ind4*, le intere scuole dove probabilmente c'è stata una diffusa interferenza degli insegnanti sullo svolgimento delle prove, considerando quelle in cui più della metà delle classi ha avuto un indicatore di *cheating* positivo.

6. Una volta eliminati dal database tutti i casi individuati dai primi quattro criteri di *cheating* si è passati all'esame dell'adattamento dei dati al modello di Rasch, individuando ed eliminando gli item che presentavano *misfit*.

7. Infine, si sono applicati gli indicatori di *misfit* relativi agli scolari, prima per classe e poi individualmente. Si sono identificate le classi con *misfit* tramite l'*Ind5* considerando come da eliminare quelle classi in cui più del 50% degli studenti presentava una condizione di *misfit*. Si è tenuta traccia dei singoli scolari che presentavano un cattivo adattamento ma che non sono stati eliminati con il quinto criterio.

Dopo aver individuato tutti i casi di *cheating* si sono ottenuti due database da cui si sono potuti selezionare gli item di link (eliminando i casi in cui *Ind1* = 1, *Ind2* = 1, *Ind3* = 1 e *Ind4* = 1) e i campioni di scolari (eliminando i casi in cui tutti gli indicatori erano uguali a 1) per la costruzione della scala di misura.

3.2 Le analisi sui test del SNV 2004/2005 e 2005/2006

In questo paragrafo si descrivono i principali risultati emersi dalle analisi dei dati dei test del SNV durante il processo di pulitura dei database. Per quanto riguarda la pulizia dei dati del test 2005/2006 si illustreranno brevemente i vari passaggi con l'ausilio di alcuni grafici, mentre per quanto riguarda il test 2004/2005 si riporteranno solamente i risultati relativi agli indicatori di *cheating* e alle stime e all'analisi del *fit* degli item secondo il modello di Rasch.

3.2.1 Analisi del *cheating*

Nelle tabelle 3.1 e 3.2 sono riportati i casi di *cheating* identificati nelle banche dati dei due test di IV elementare, suddivisi nelle cinque macro aree geografiche.

I grafici 3.6 - 3.10 rappresentano, invece, le percentuali di casi di *cheating* rilevati con i cinque criteri nelle province di ciascuna macro area geografica. È facile notare le differenze che intercorrono tra una macro area e l'altra; per esempio, tra il Nord Est (l'area dove è stato individuato il numero minore di casi anomali) e il Sud e Isole (l'area che ha presentato il numero maggiore):

Macroarea	Pop. origin.	Ind1	Ind2	Ind3	Ind4	Ind5	Pop. finale
SNV 2004/2005							
Nord Ovest	42.638	35	1.591	26	187	102	40.697
Nord Est	151.727	78	2.619	71	168	198	148.593
Centro	86.177	242	5.712	44	992	206	78.981
Sud	114.185	1.221	23.763	328	5.716	299	82.858
Sud e isole	86.170	972	17.415	230	3.849	233	63.471
Totale	480.897	2.548	51.100	699	10.912	1.038	414.600

Tabella 3.1: Casi di *cheating* o malfunzionamento nel test 2004/2005

Macroarea	Pop. origin.	Ind1	Ind2	Ind3	Ind4	Ind5	Pop. finale
SNV 2005/2006							
Nord Ovest	43.801	54	2.072	52	298	207	41.118
Nord Est	157.267	81	3.112	77	172	256	153.569
Centro	88.493	330	5.556	40	837	560	81.170
Sud	116.243	1.307	23.697	538	5.299	1.032	84.370
Sud e isole	85.808	1.036	18.504	434	4.155	963	60.716
Totale	491.612	2.808	52.941	1.141	10.761	3.018	420.943

Tabella 3.2: Casi di *cheating* o malfunzionamento nel test 2005/2006

nelle prime le percentuali sono tutte al di sotto del 10%, nelle seconde la maggior parte supera il 20%, con alcune che raggiungono addirittura il 30%.

Le tabelle 3.3 e 3.4 riportano le stime delle difficoltà di tutti gli item (ottenute con i due software RUMM2020 e Winsteps) nei due campioni (*Camp1* e *Camp2*) dopo aver eliminato dalla popolazione solo i soggetti positivi al primo criterio di *cheating* (i casi individuati da $Ind1 = 1$), sull'intera popolazione e sulla popolazione pulita (avendo eliminato tutti i casi che presentavano almeno un indicatore di *cheating* uguale a 1). In questa analisi il confronto viene fatto su tutti gli item, anche se, come si vedrà nel prossimo paragrafo, 3 item nel test 2004/2005 (l'item 9, l'item 11 e l'item 20) e 3 item nel test 2005/2006 (l'item 6, l'item 10 e l'item 16) hanno rivelato un elevato *misfit* e sono stati scartati nel momento in cui si è andati a definire il gruppo di domande per la costruzione del test di link e la scala di misurazione.
Si può affermare che, in generale, non ci sono differenze apprezzabili nelle difficoltà degli item da un campione all'altro, nè tra quelle stimate sui campioni

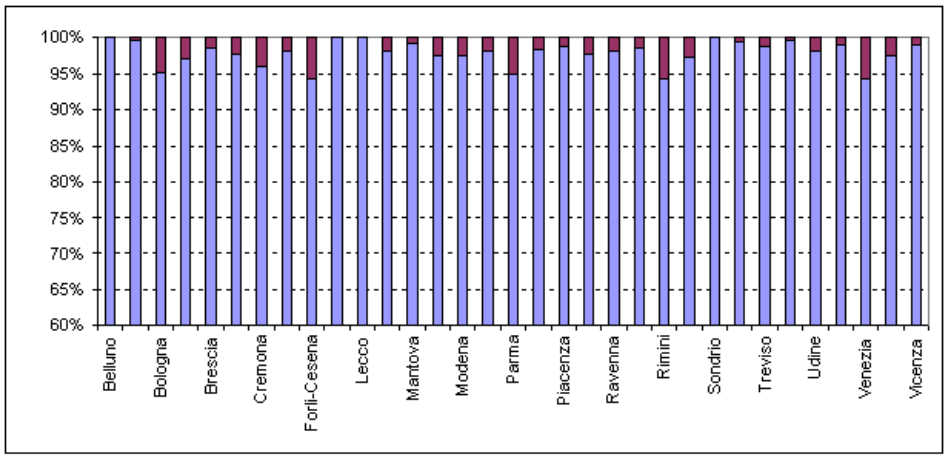


Figura 3.6: Percentuale dei casi eliminati nelle province del Nord Est (SNV 2005/2006)

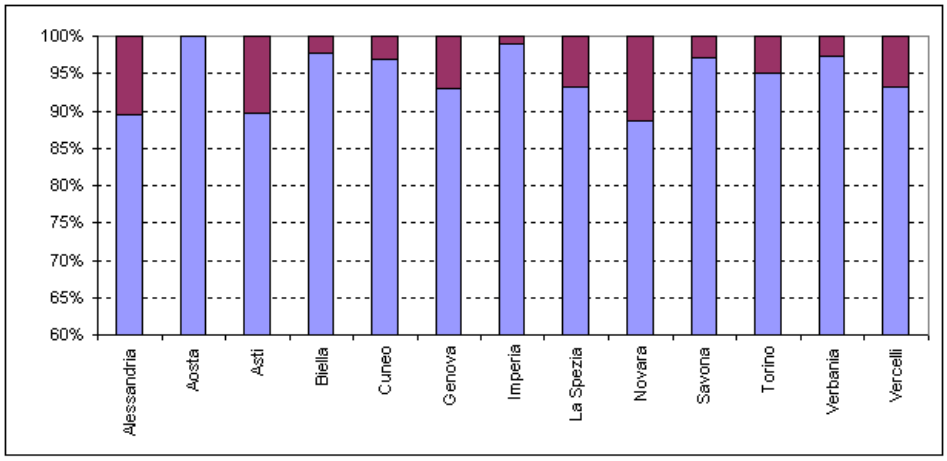


Figura 3.7: Percentuale dei casi eliminati nelle province del Nord Ovest (SNV 2005/2006)

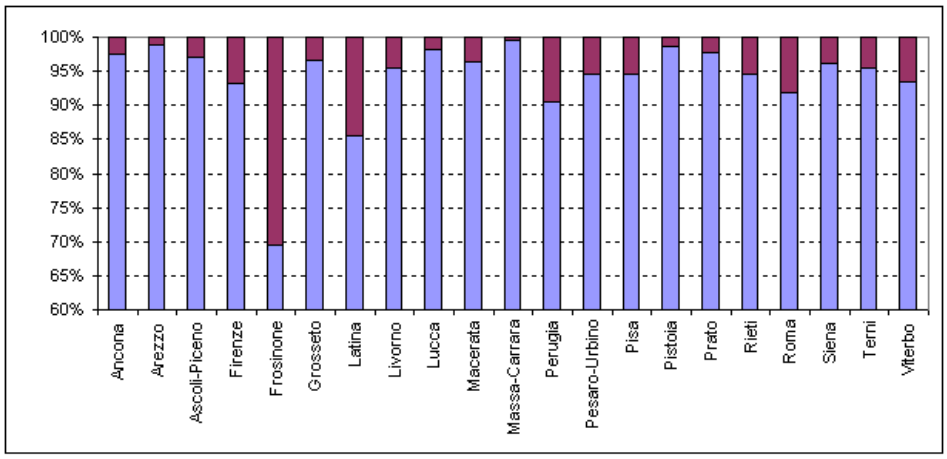


Figura 3.8: Percentuale dei casi eliminati nelle province del Centro Italia (SNV 2005/2006)

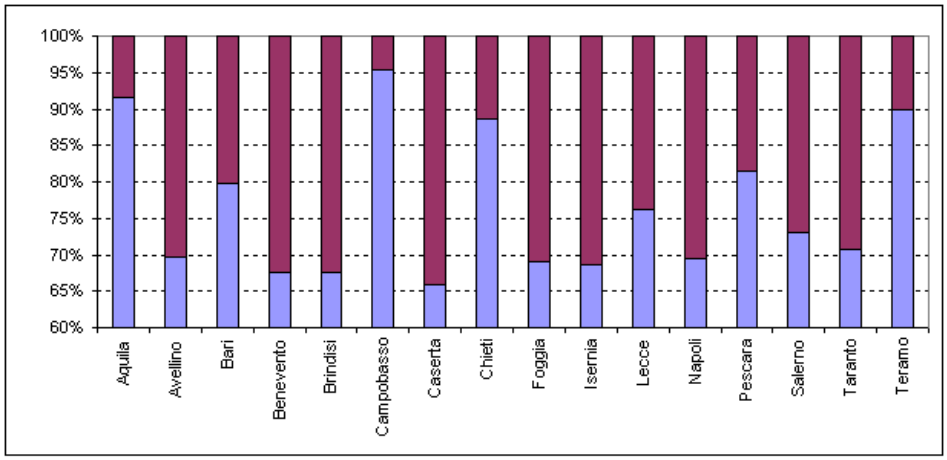


Figura 3.9: Percentuale dei casi eliminati nelle province del Sud Italia (SNV 2005/2006)

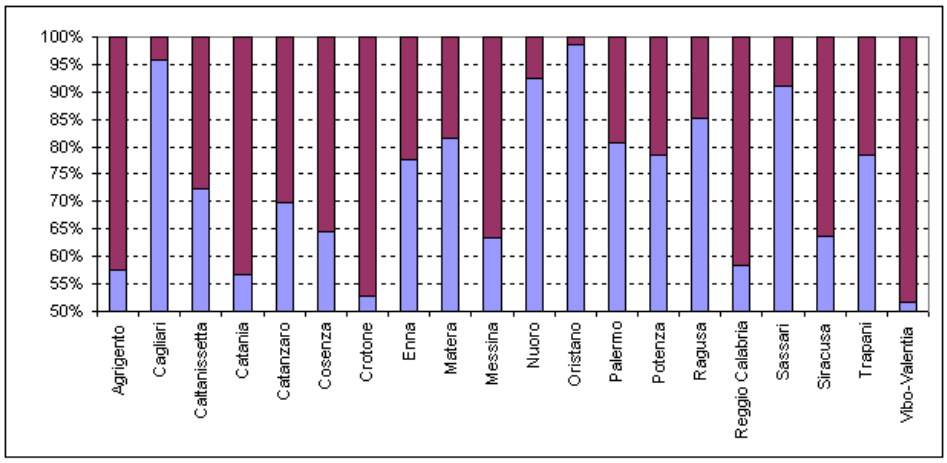


Figura 3.10: Percentuale dei casi eliminati nelle province del Sud e Isole (SNV 2005/2006)

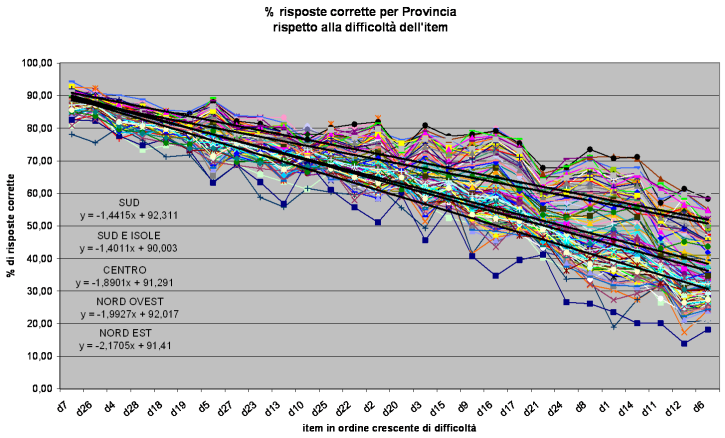


Figura 3.11: Percentuali risposte corrette per ogni item (SNV 2005/2006; database originale)

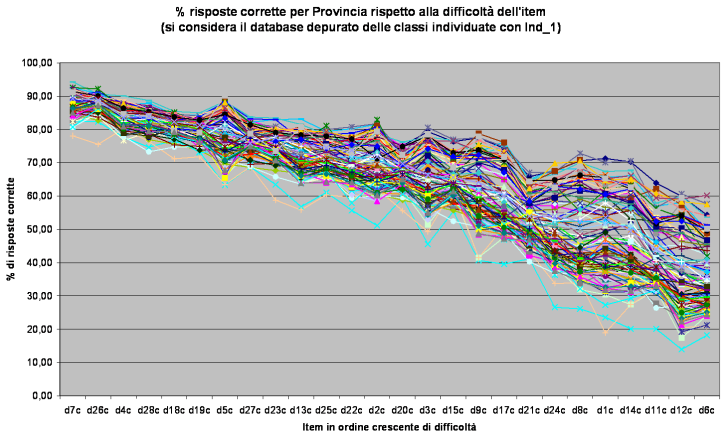


Figura 3.12: Percentuali risposte corrette per ogni item (SNV 2005/2006: Ind1=0)

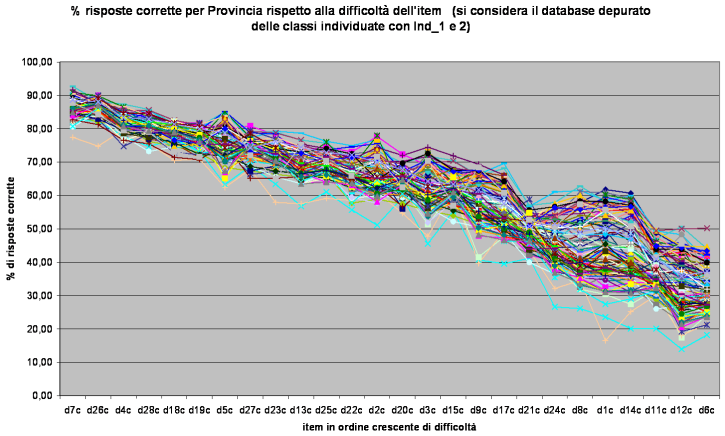


Figura 3.13: Percentuali risposte corrette per ogni item (SNV 2005/2006: Ind1=0 e Ind2=0)

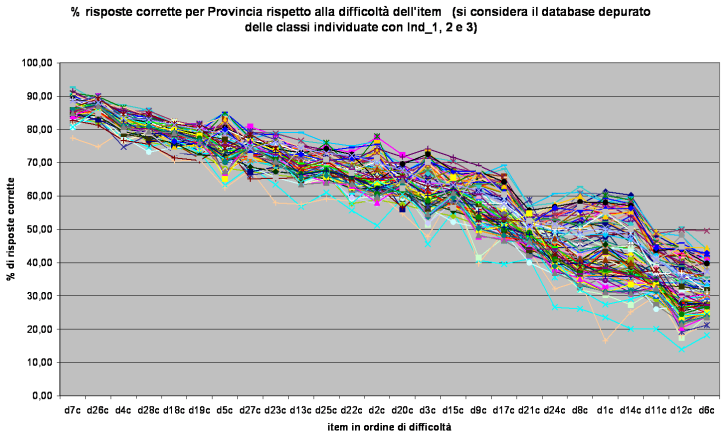


Figura 3.14: Percentuali risposte corrette per ogni item (SNV 2005/2006: Ind1=0, Ind2=0 e Ind3=0)

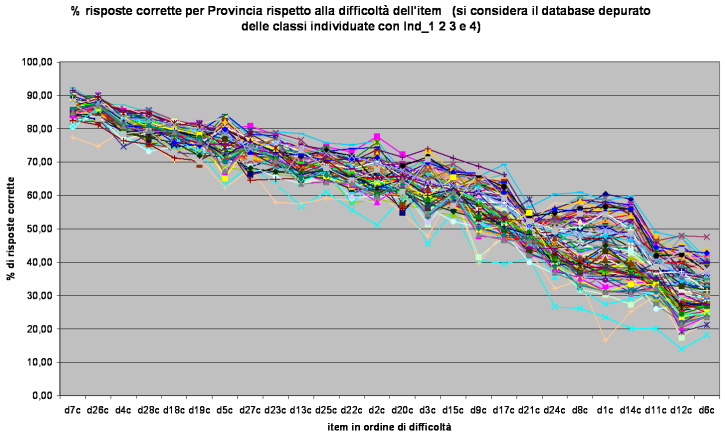


Figura 3.15: Percentuali risposte corrette per ogni item (SNV 2005/2006: Ind1=0, Ind2=0, Ind3=0 e Ind4=0)

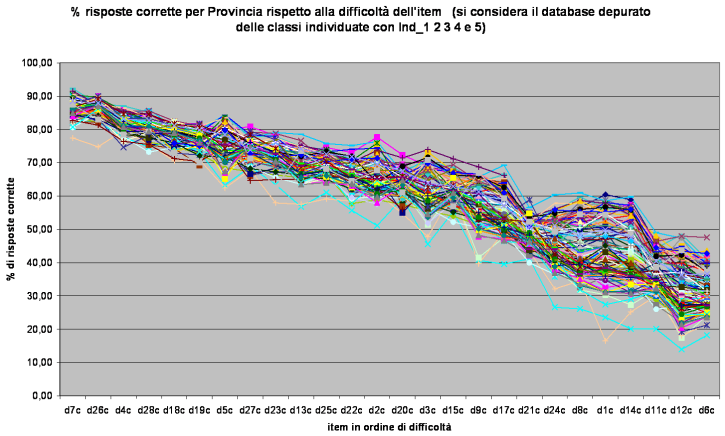


Figura 3.16: Percentuali risposte corrette per ogni item (SNV 2005/2006: Ind1=0, Ind2=0, Ind3=0, Ind4=0 e Ind5=0)

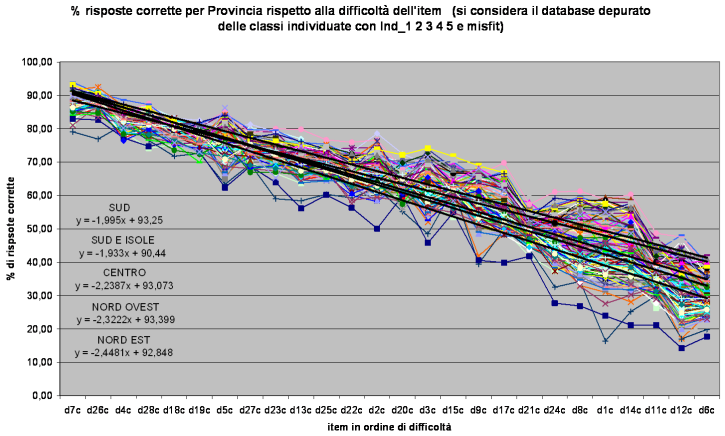


Figura 3.17: Percentuali risposte corrette per ogni item (SNV 2005/2006: Ind1=0, Ind2=0, Ind3=0, Ind4=0, Ind5=0 e senza misfit)

e quelle stimate sull'intera popolazione.

I grafici 3.11 - 3.17, poi, rappresentano, per ogni provincia, le percentuali di risposte corrette a ciascun item, ordinati dal più facile al più difficile. Il grafico 3.11 si riferisce al database originale completo di tutti i dati; i grafici successivi sono ottenuti dal database depurato, di volta in volta, di tutti i casi positivi a ciascuno dei cinque criteri di *cheating* definiti precedentemente. Le province che, in origine, avevano un andamento poco inclinato a causa dell'alta percentuale di risposte corrette agli item più difficili, mano a mano che si eliminano i casi di *cheating*, cominciano ad assumere un andamento più plausibile e simile a quelle delle province dove il *cheating* ha avuto meno peso; così le linee interpolatrici dei grafici 3.11 e 3.17, che rappresentano le rette di regressione nelle cinque macro aree, si avvicinano l'una all'altra, pur rimanendo distinte.

3.2.2 Analisi degli item

Durante il processo di pulizia dei database è stata condotta anche un'analisi sugli item dei questionari per verificare il loro adattamento al modello di Rasch. Di seguito si riportano i principali risultati.

Le figure 3.18 e 3.19 rappresentano le *Item-Person Map* dei due test determinate sulle popolazioni originarie con tutti i 28 item presenti nei questionari. La linea retta verticale rappresenta il *continuum*, cioè la scala a intervalli su cui si collocano le stime delle difficoltà degli item (a destra, dal più facile, in basso, al più difficile, in alto) e le stime delle abilità degli individui (a sinistra, dal meno capace, in basso, al più capace, in alto). L'origine della scala è fissata arbitrariamente sulla media delle difficoltà degli item, l'unità di misura comune alle persone e agli item è il *logit*; dato che la scala di misura è una scala a intervalli, l'unità di misura è costante, da cui ne consegue che distanze uguali in punti differenti del *continuum* misurano la stessa variazione di intensità e i confronti, fra item e individui, possono essere effettuati semplicemente sottraendo tra loro i valori delle loro difficoltà o abilità.

item	c1-R	item	c1-W	item	c2-R	item	c2-W	item	tutti-W	item	puliti-W
20	1,49	20	1,80	20	1,59	20	1,77	20	1,80	20	1,71
22	1,24	22	1,37	22	1,27	22	1,35	22	1,37	22	1,38
21	1,08	21	1,13	21	1,10	21	1,10	21	1,14	21	1,17
17	0,82	17	0,83	17	0,83	16	0,81	17	0,83	17	0,87
19	0,68	16	0,80	16	0,75	17	0,80	16	0,80	16	0,76
16	0,68	19	0,68	13	0,65	19	0,68	19	0,67	19	0,71
13	0,61	13	0,65	19	0,64	13	0,65	13	0,65	13	0,67
27	0,59	27	0,64	27	0,60	27	0,64	27	0,64	27	0,61
9	0,57	9	0,46	9	0,48	9	0,46	9	0,46	9	0,50
11	0,47	11	0,44	11	0,41	11	0,40	11	0,46	11	0,48
10	0,06	5	0,07	18	0,05	5	0,05	18	0,06	18	0,07
18	0,06	18	0,05	5	0,01	18	0,05	5	0,05	5	0,02
5	0,03	10	-0,04	10	-0,04	10	-0,04	10	-0,03	10	-0,01
25	-0,03	25	-0,08	25	-0,07	25	-0,08	25	-0,08	25	-0,07
2	-0,17	2	-0,16	2	-0,14	2	-0,16	2	-0,14	2	-0,15
24	-0,25	23	-0,21	23	-0,19	23	-0,21	23	-0,21	23	-0,20
23	-0,25	24	-0,25	24	-0,23	24	-0,25	24	-0,26	24	-0,27
3	-0,38	3	-0,35	3	-0,35	3	-0,34	3	-0,35	3	-0,34
28	-0,42	28	-0,44	28	-0,41	28	-0,40	28	-0,46	28	-0,48
4	-0,48	4	-0,47	4	-0,48	4	-0,46	4	-0,50	4	-0,49
7	-0,61	7	-0,67	7	-0,64	6	-0,68	7	-0,66	7	-0,66
6	-0,64	6	-0,68	6	-0,65	7	-0,72	6	-0,67	6	-0,68
26	-0,68	26	-0,75	26	-0,73	26	-0,75	26	-0,74	26	-0,75
14	-0,70	14	-0,79	14	-0,77	14	-0,79	14	-0,77	14	-0,77
1	-0,79	1	-0,86	1	-0,82	1	-0,86	1	-0,87	1	-0,87
15	-0,89	15	-0,95	15	-0,91	15	-0,95	15	-0,95	15	-0,97
8	-1,03	12	-1,08	12	-1,00	12	-1,02	12	-1,08	12	-1,12
12	-1,06	8	-1,12	8	-1,05	8	-1,07	8	-1,14	8	-1,15

Tabella 3.3: Misure degli item nel 2004/2005

item	c1-R	item	c1-W	item	c2-R	item	c2-W	item	tutti-W	item	puliti-W
6	1,57	6	1,74	6	1,55	6	1,72	6	1,74	6	1,76
12	1,53	12	1,65	12	1,53	12	1,65	12	1,66	12	1,73
11	1,32	11	1,41	11	1,32	11	1,41	11	1,41	11	1,43
14	1,05	14	1,08	14	1,06	14	1,09	14	1,08	14	1,1
1	0,91	1	1,00	1	0,91	1	1,00	1	1,01	1	1,02
8	0,89	8	0,95	8	0,90	8	0,95	8	0,95	8	1
24	0,85	24	0,86	24	0,87	24	0,88	24	0,88	24	0,91
21	0,75	21	0,75	21	0,74	21	0,74	21	0,74	21	0,73
17	0,40	17	0,40	17	0,38	17	0,38	17	0,4	17	0,42
16	0,26	9	0,27	16	0,26	9	0,27	9	0,26	16	0,3
9	0,25	16	0,27	9	0,25	16	0,27	16	0,26	9	0,28
15	0,12	15	0,11	15	0,10	15	0,09	15	0,1	15	0,08
3	0,05	3	0,05	3	0,05	3	0,05	3	0,04	3	0,05
20	0,02	20	0,00	20	0,04	20	0,02	20	0	20	-0,03
2	-0,14	2	-0,14	2	-0,14	2	-0,15	2	-0,15	2	-0,14
22	-0,15	22	-0,17	22	-0,15	22	-0,16	22	-0,16	22	-0,17
25	-0,28	25	-0,33	10	-0,26	10	-0,33	10	-0,33	13	-0,37
10	-0,28	10	-0,34	25	-0,29	25	-0,34	25	-0,33	25	-0,37
13	-0,34	13	-0,36	13	-0,34	13	-0,36	13	-0,36	10	-0,42
23	-0,52	23	-0,55	23	-0,51	23	-0,55	23	-0,53	23	-0,54
27	-0,55	27	-0,59	27	-0,57	27	-0,62	27	-0,63	27	-0,64
5	-0,68	5	-0,71	5	-0,68	5	-0,71	5	-0,72	5	-0,71
19	-0,86	19	-0,90	19	-0,86	19	-0,89	19	-0,89	19	-0,9
18	-0,93	18	-0,97	18	-0,92	18	-0,97	18	-0,95	18	-0,98
28	-1,02	28	-1,08	28	-1,02	28	-1,09	28	-1,09	28	-1,12
4	-1,14	4	-1,18	4	-1,16	4	-1,21	4	-1,2	4	-1,21
26	-1,55	7	-1,61	7	-1,52	7	-1,57	26	-1,59	7	-1,61
7	-1,56	26	-1,61	26	-1,52	26	-1,58	7	-1,59	26	-1,63

Tabella 3.4: Misure degli item nel 2005/2006

Ad esempio, l'item 3 del test 2005/2006, posto, grosso modo, sull'origine della scala, è più facile dell'item 1 quanto l'item 18 è più facile dell'item 3. Il medesimo principio di valore uguale dell'unità di misura della scala si applica naturalmente alle differenze tra abilità degli individui. Nel modello di Rasch queste ultime sono definite in modo che ciascun individuo abbia il 50% di probabilità di superare un item di difficoltà pari alla sua abilità, quindi un item la cui difficoltà si trova nello stesso punto del suo parametro stimato sulla scala. Ad esempio, una persona con un'abilità pari a 0 ha il 50% di probabilità di rispondere correttamente all'item 3. Questa stessa persona dovrebbe avere più del 50% di probabilità di superare con successo item più facili e meno del 50% di probabilità di superare con successo item più difficili. Quando le risposte osservate sono molto differenti da quanto ci si aspetterebbe, in termini di valori attesi, si è in presenza di un *misfit* di persona (*person misfit*).

Osservando le due *Item-Person Map* si nota che in entrambe le prove del SNV la distribuzione delle abilità degli scolari non è centrata sulla distribuzione degli item ma è spostata verso l'alto (la differenza tra la media delle persone e la media degli item è di circa 1.15 *logit* nel 2004/2005 e 0.75 nel 2005/2006), ha una coda destra molto allungata che raggiunge il +4 sulla scala e una buona parte degli individui si trova al di sopra degli item più difficili (item 20, item 21 e item 22 nel 2004/2005 e item 6, item 11 e item 12 nel 2005/2006). Da questo si deduce che i questionari delle prove sono stati troppo facili per le coorti di studenti per cui erano stati predisposti; il fatto poi, che moltissimi scolari hanno ottenuto *perfect scores* con 28 risposte corrette su 28 domande è un indizio in più che avvalora l'ipotesi che l'intervento esterno degli insegnanti ha, almeno in parte, falsato i risultati della valutazione.

Si è già osservato precedentemente che la pulitura del database dai casi di *cheating* non ha avuto un effetto significativo sulle stime delle difficoltà degli item nonostante le percentuali di risposte corrette agli item più impegnativi si siano abbassate; le *Item-Person Map* costruite sulle due coorti, dopo aver eliminato i record con *Ind1*, *Ind2*, *Ind3*, *Ind4* e *Ind5* uguali a 0, confermano che con la pulitura non si sono registrate apprezzabili modifiche della scala ma che, invece, sono cambiate le distribuzioni delle abilità, che hanno assunto una forma molto più simmetrica rispetto a quella originaria. Tutte le analisi e

i risultati che seguono sono condizionati dal fatto che entrambe le prove non erano calibrate correttamente sulle abilità da valutare; sarebbero necessarie prove con item più difficili per poter avere misure più attendibili per i livelli più alti di conoscenza e competenza della matematica.

Un altro aspetto importante inerente alla validità dello strumento di misurazione riguarda la "bontà" delle singole domande del questionario. Nella teoria di Rasch, fondata sui rigidi principi di monotonicità, unidimensionalità e indipendenza locale, l'analisi degli item svolge un ruolo centrale in quanto ogni item, di concerto con tutti gli altri, concorre a determinare la stima del tratto latente da misurare. In questa ricerca, per valutare l'adattamento degli item al RM si è fatto ricorso alle statistiche basate sui punteggi totali: l'*Outfit* e l'*Infit* espressi, rispettivamente, dalle 2.67 e 2.69 e presenti in Winsteps e l'*Item Trait Interaction Test of Fit* espresso dalla 2.72 e presente in RUMM2020.

In Winsteps queste statistiche sono riportate come medie dei quadrati dei residui nella forma di un chi-quadrato diviso per i suoi gradi di libertà, in modo da assumere sempre valori positivi e, approssimativamente, la distribuzione di una v.c. di valore atteso pari a +1 e un range che varia da 0 a $+\infty$. Un valore di *Infit* o *Outfit mean square* uguale a $1 + x$ indica il 100x% di variazione in più (se x è positivo) o in meno (se x è negativo) di quanto ci si sarebbe aspettati se i dati si adattassero perfettamente al modello, tra i comportamenti di risposta osservati e quelli predetti dal modello. Quindi un *Outfit mean square value* pari a 1,30 (1+0,30) indica il 30% di variazione in più nei dati osservati di quella predetta dal RM; un *Outfit mean square value* minore di 1, ad esempio pari a 0,78 (1-0,22=0,78) indica il 22% di variazione in meno nei dati osservati di quella predetta dal modello. Considerare l'*Infit* o l'*Outfit* significa dar più importanza agli item di difficoltà prossima all'abilità del soggetto oppure a quelli di difficoltà più distante dalla sua abilità.
L'idea di *misfit* della stringa di risposta che mostra più variazione di quanto ci si aspetti è un concetto intuitivo e riconosciuto universalmente nelle analisi di adattamento. Con riferimento all'abilità di una persona ciò accade quando le risposte osservate sono più discordanti di quelle attese per l'effetto dell'errore casuale; si riscontrano risposte sbagliate ad item semplici oppure risposte corrette ad item difficili. Meno immediato è il concetto di *misfit* dei casi che

mostrano meno variabilità di quella attesa. Esso si rifà al concetto di *response independence* in base al quale la probabilità di rispondere ad ogni item deve dipendere solo dall'abilità del soggetto e non da come questi ha riposto agli altri item. Violazioni della *response independence* si manifestano con *pattern* di risposta che richiamano una struttura deterministica e *mean square residual* piccoli.

Le tabelle 3.20 e 3.21 rappresenta l'output principale di RUMM2020 per quanto riguarda le statistiche sull'analisi del *fit* degli item; in entrambi i casi la numerosità delle osservazioni per il calcolo del *fit* è stata diminuita per ridurre l'effetto inflattivo sul chi-quadrato (vedi paragrafo 2.4.1). Nell'ordine compaiono: il numero progressivo dell'item (*Item*), la difficoltà stimata (*Location*), il suo standard error (*SE*), il residuo standardizzato (*Residual*), i suoi gradi di libertà (*DF*), la statistica *Item Trait Interaction Test of Fit* (*ChiSq*), i suoi gradi di libertà (*DF*) e il suo *p-value* (*Prob*).

L'item 6, l'item 10 e l'item 16 del test 2005/2006 hanno un *misfit* significativo ed evidenziano scarso adattamento al modello. La controparte grafica dell'analisi del *fit* è costituita dalle *ICC*. Nella fattispecie le *ICC* delle domande 6, 10 e 16 confermano il cattivo adattamento di questi 3 item. La figura 3.22 rappresenta l'ICC dell'item 6; si nota che le percentuali di risposte corrette osservate (i puntini neri) hanno un andamento molto piatto per valori di abilità attorno allo 0, poi aumentano rimanendo al di sotto della curva teorica. Trattandosi di un item molto difficile è probabile che in questo caso il fenomeno del *teacher cheating* ne abbia falsato il comportamento. L'alta percentuale di risposte corrette per le abilità più basse non è plausibile; immaginando di considerare solamente le altre tre percentuali, l'andamento della curva sarebbe conforme a un'ogiva logistica, in questo caso relativa a un item con difficoltà stimata ancora più elevata di quella stimata.

Nella figura 3.23 l'*ICC* dell'item 10 ha le probabilità osservate più alte di quelle attese per le abilità più piccole e più basse di quelle attese per le abilità più grandi. È un tipico esempio in cui l'item non discrimina a sufficienza tra le abilità degli individui (*under discrimination*) e la causa è spesso legata a una violazione dell'ipotesi di unidimensionalità (l'item va a cogliere, oltre alla variabile oggetto della misurazione, anche un'altra variabile).

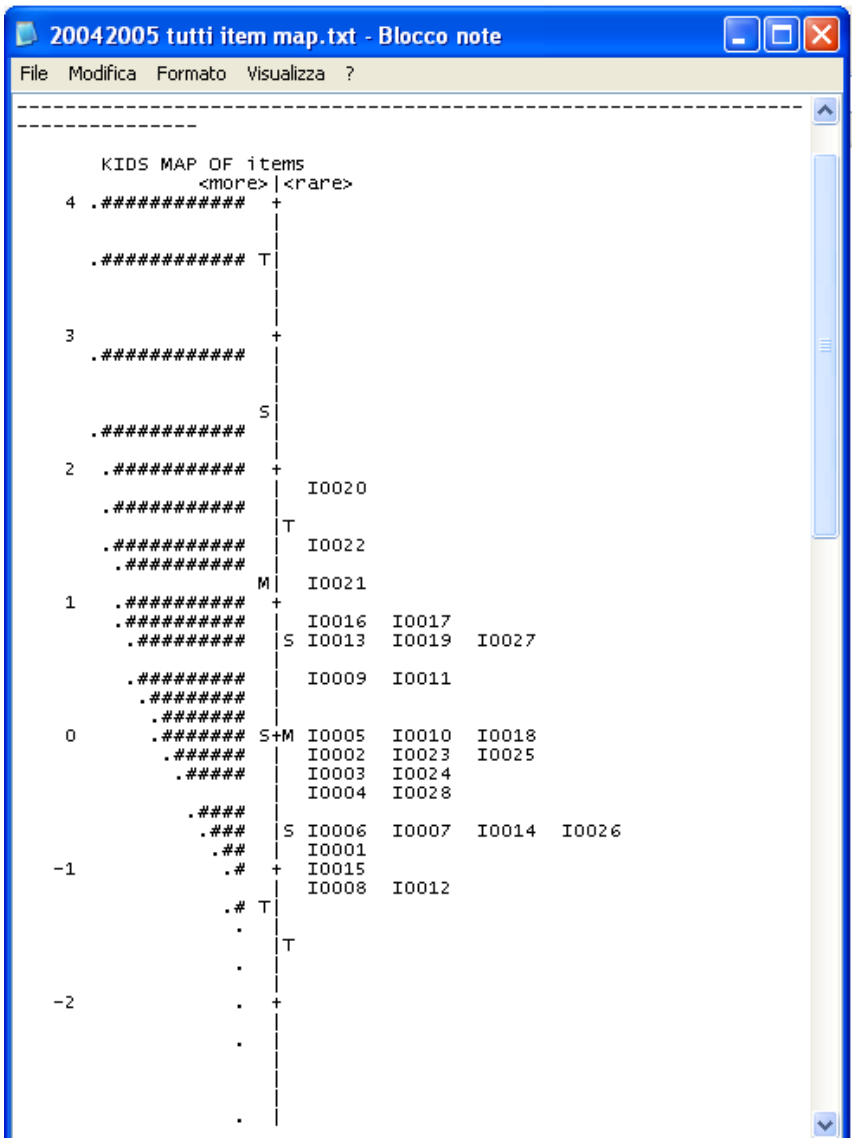


Figura 3.18: Item-Person Map del test 2004/2005

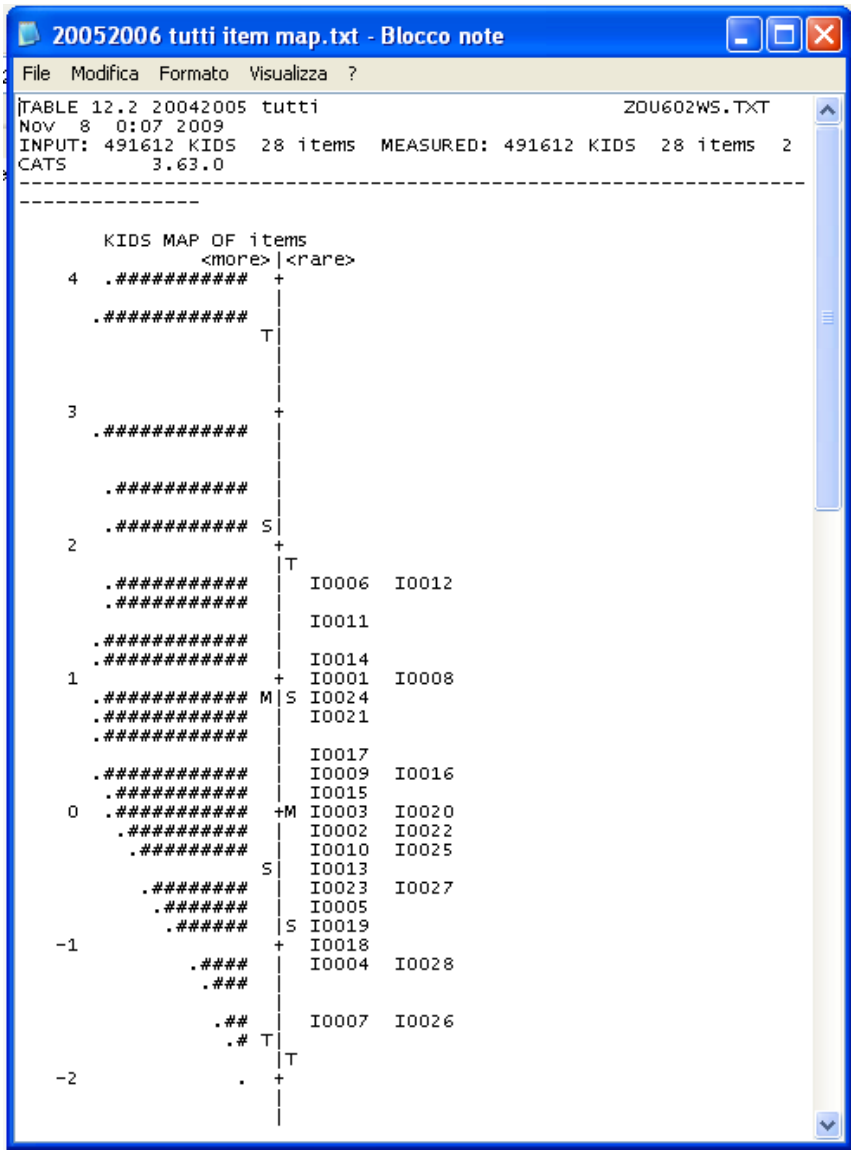


Figura 3.19: Item-Person Map del test 2005/2006

Seq	Location	SE	Residual	DF	ChiSq	DF	Prob
1	-0,821	0,013	-3,987	4033,11	1,999	3	0,572681
2	-0,139	0,011	2,226	4033,11	1,229	3	0,746169
3	-0,348	0,012	-15,025	4033,11	10,159	3	0,017261
4	-0,467	0,012	-17,768	4033,11	17,980	3	0,000445
5	0,079	0,011	-0,591	4033,11	1,635	3	0,651408
6	-0,648	0,013	-8,041	4033,11	4,155	3	0,245160
7	-0,635	0,013	-4,628	4033,11	1,263	3	0,737956
8	-1,047	0,014	0,855	4033,11	1,243	3	0,742784
9	0,481	0,011	-34,468	4033,11	33,914	3	0,000000
10	-0,035	0,011	-11,780	4033,11	6,212	3	0,101754
11	0,411	0,011	16,277	4033,11	14,442	3	0,002363
12	-1,000	0,014	6,075	4033,11	3,855	3	0,277507
13	0,642	0,011	-27,304	4033,11	24,751	3	0,062019
14	-0,767	0,013	-9,080	4033,11	2,785	3	0,425982
15	-0,912	0,013	-3,830	4033,11	1,679	3	0,641696
16	0,751	0,010	8,638	4033,11	1,805	3	0,613911
17	0,827	0,010	-26,725	4033,11	23,885	3	0,000028
18	0,052	0,011	11,916	4033,11	9,078	3	0,028269
19	0,636	0,011	7,132	4033,11	3,506	3	0,320039
20	1,594	0,011	28,085	4033,11	21,019	3	0,000105
21	1,101	0,010	-8,559	4033,11	7,152	3	0,067202
22	1,272	0,010	10,537	4033,11	2,837	3	0,417502
23	-0,191	0,012	-2,482	4033,11	1,570	3	0,666306
24	-0,230	0,012	-13,219	4033,11	8,481	3	0,037052
25	-0,071	0,011	1,357	4033,11	3,175	3	0,365417
26	-0,731	0,013	-9,668	4033,11	4,511	3	0,211356
27	0,604	0,011	13,919	4033,11	3,859	3	0,277151
28	-0,410	0,012	-3,992	4033,11	1,938	3	0,585378

Figura 3.20: Individual Item Fit del test 2004/2005

Item	Loc	SE	Residual	DF	ChiSq	DF	Prob
01	0,958	0,035	4,691	4002,75	4,581	3	0,205167
02	-0,160	0,036	0,639	4002,75	0,168	3	0,982601
03	0,070	0,036	-1,395	4002,75	0,736	3	0,864768
04	-1,105	0,043	-0,660	4002,75	2,381	3	0,497267
05	-0,633	0,039	-1,791	4002,75	0,680	3	0,877832
06	1,563	0,037	6,827	4002,75	14,841	3	0,001958
07	-1,645	0,049	-0,768	4002,75	1,150	3	0,764900
08	0,917	0,035	-0,364	4002,75	4,070	3	0,254002
09	0,265	0,035	1,228	4002,75	1,419	3	0,701184
10	-0,384	0,037	6,723	4002,75	7,902	3	0,048092
11	1,311	0,036	0,721	4002,75	0,657	3	0,883176
12	1,552	0,037	3,289	4002,75	3,505	3	0,320059
13	-0,340	0,037	-1,404	4002,75	1,200	3	0,753012
14	1,069	0,035	-3,569	4002,75	7,609	3	0,054828
15	0,072	0,036	3,024	4002,75	0,644	3	0,886375
16	0,326	0,035	-9,005	4002,75	17,798	3	0,000485
17	0,431	0,035	-2,904	4002,75	3,948	3	0,267161
18	-0,960	0,041	-2,251	4002,75	4,235	3	0,237173
19	-0,819	0,040	-1,684	4002,75	2,715	3	0,437722
20	-0,012	0,036	0,925	4002,75	0,182	3	0,980513
21	0,674	0,035	-2,537	4002,75	4,239	3	0,236767
22	-0,143	0,036	2,704	4002,75	1,062	3	0,786331
23	-0,563	0,038	-2,507	4002,75	3,696	3	0,296258
24	0,898	0,035	-1,531	4002,75	2,092	3	0,553609
25	-0,330	0,037	4,026	4002,75	2,415	3	0,490917
26	-1,489	0,047	-0,044	4002,75	0,908	3	0,823491
27	-0,548	0,038	-2,674	4002,75	2,170	3	0,537813
28	-0,975	0,041	0,892	4002,75	2,888	3	0,409197

Figura 3.21: Individual Item Fit del test 2005/2006

Nella figura 3.24, che rappresenta l'*ICC* dell'item 16, invece, si nota un fenomeno opposto: i valori osservati si dispongono in modo più inclinato di quanto richiederebbe il modello. È un caso tipico di *over discrimination*, cioè un caso in cui l'item discrimina troppo tra le abilità degli individui (piccole variazioni di abilità producono forti variazioni nella frequenza delle risposte corrette), legato spesso a un problema di *response dependence*, per cui le risposte all'item dipendono anche da come un soggetto ha risposto ad altri quesiti. Se si va a vedere il test del 2005/2006 si può notare che l'item 16 è molto simile, nel contenuto, all'item 14 e per questo ridondante; entrambi richiedono la risoluzione di un'uguaglianza ed è facile supporre che chi ha risposto correttamente all'item 14 abbia risposto correttamente anche all'item 16 e chi non ha saputo rispondere alla prima domanda non abbia saputo rispondere neppure alla seconda.

Si indicano a titolo esemplificativo due degli item eliminati: il 10 e il 16.

Esempio

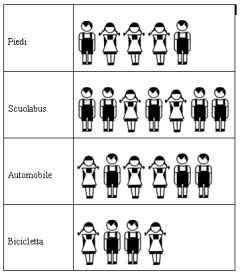
Domanda 10. Il grafico rappresenta con quale mezzo di trasporto sono andati a scuola oggi le bambine e i bambini di una classe. Quanti bambini maschi hanno usato lo scuolabus o l'automobile?

A. 3

B. 5

C. 8

D. 13



Esempio

Domanda 16. Quale valore deve avere il ▲ perché l'uguaglianza sia vera?

$$2 * 10 = 20 * \blacktriangle$$

A. 1
B. 10
C. 100
D. 1000

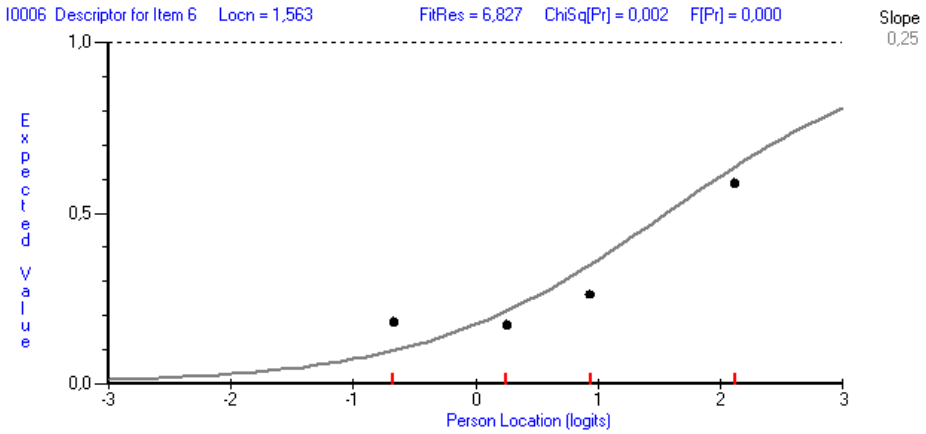


Figura 3.22: ICC dell'item 6 del test 2005/2006

Anche nel test del SNV del 2004/2005 gli item che si adattano meno al modello sono tre: l'item 9 (*over discrimination*), l'item 11 (*under discrimination*) e l'item 20 (*under discrimination*).

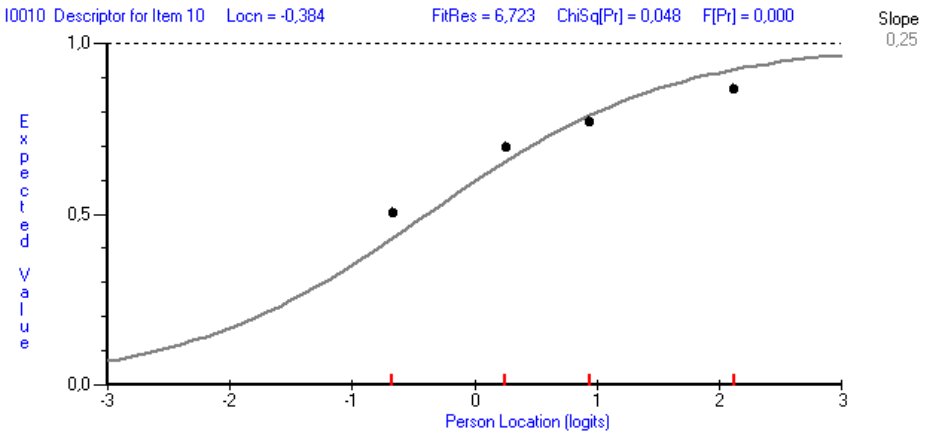


Figura 3.23: ICC dell'item 10 del test 2005/2006

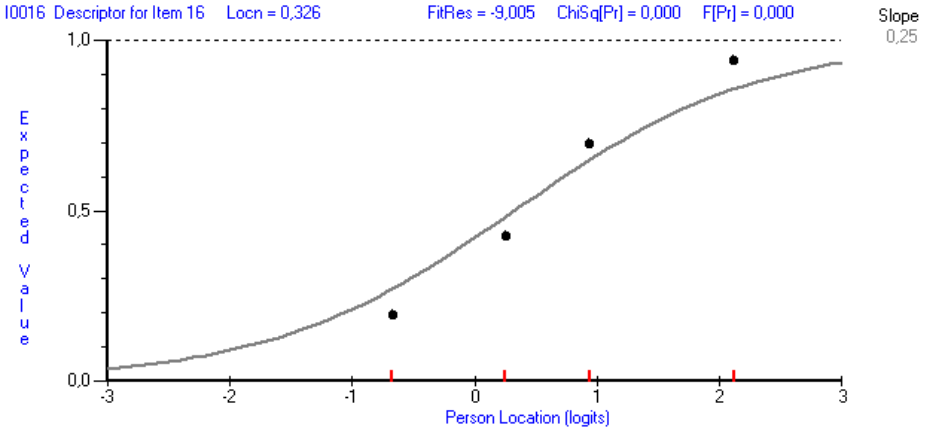


Figura 3.24: ICC dell'item 16 del test 2005/2006

Lo studio sul comportamento degli item ha riguardato anche l'analisi del *DIF*. I diversi casi di *DIF* rilevati sia nel test del 2004/2005 che nel test del 2005/2006 sono stati esclusi dal gruppo di item da cui sono state scelte le domande di aggancio per legare assieme le tre prove (item 1, 22, 23, 25 e 27 del SNV 2004/2005 e item 9, 11, 13, 15, 21, 24 e 28 del SNV 2005/2006), ma sono stati mantenuti nella fase di costruzione della scala di misurazione. Questo per due motivi: a) per avere un numero di item sufficientemente adeguato per misurare le abilità nei due anni e b) per l'impossibilità di distinguere i casi di *DIF* reale dai casi di *DIF* apparente o dovuto ad altri fattori (per esempio che il fenomeno del *cheating* è stato più rilevante in alcune province che in altre). Qui di seguito si riportano quattro esempi (sempre relativi agli item del 2005/2006) di studio del *bias* con i risultati dell'ANOVA dei residui.

L'item 7 (fig. 3.26 e 3.27) evidenzia un buon *fit* complessivo e assenza di *DIF* tra le cinque macro aree considerate: l'effetto classe intervallo (*ANOVA-FIT[CInt]*, *p-value* = 0,483), l'effetto macro area (*DIF[area]*, *p-value* = 0,479) e l'effetto congiunto tra la classe intervallo e la macro area (*area-by-Cint*, *p-value* = 0,999) non sono significativi.

L'item 6 (fig. 3.28 e 3.29) non rivela un effetto dovuto all'area geografica (l'*F-ratio* relativo al *DIF [area]* è 0,73 con una probabilità pari a 0,57) ma un adattamento dovuto all'effetto classe (*Anova-Fit[CInt]*) molto basso (*F-ratio* = 14,04 con *p-value* prossimo allo 0). Osservando la figura 3.28 si nota che le cinque spezzate che rappresentano le percentuali osservate di risposte corrette nelle cinque macro aree sono vicine le une alle altre, ma tutte si discostano sensibilmente dalla curva teorica, soprattutto per i livelli di abilità bassi. Il *misfit* totale di conseguenza risulta significativo (7,86, per un *F-ratio* pari a 2,48 e un *p-value* pari a 0,000), come già rilevato in precedenza, mentre l'effetto *DIF* complessivo (*Total Item DIF*) è assente (*F-ratio* = 0,30 con *p-value* = 0,996).

Al contrario, come si osserva dalle figure 3.30 e 3.31, l'item 27 ha un effetto *DIF* abbastanza marcato (*p-value* pari a 0,019 e le spezzate che rappresentano le probabilità osservate ben separate tra loro, soprattutto quelle del Centro, del Sud e del Sud e Isole da quelle del Nord Est e del Nord Ovest) ma non presenta malfunzionamento per quanto riguarda il solo effetto della *CI* (*p-value* pari a

0,156) e l'effetto combinato *area-by-CInt* (*p-value* pari a 0,942). In questo caso il *misfit* totale è imputabile al fatto che l'item si comporta in maniera differente nelle cinque macro aree; il fatto, poi, che le spezzate siano grosso modo parallele indica che si è in presenza di un *DIF* uniforme dovuto alla differente difficoltà dell'item stimata nelle diverse regioni (l'item è risultato più facile al Nord che al Centro e al Sud).
Nell'item 16 (fig. 3.32 e 3.33), infine, risultano molto significativi sia l'effetto classe intervallo (*p-value* prossimo a zero) che l'effetto del *DIF* per macro aree (*p-value* = 0,005); di conseguenza anche il *misfit* totale risulta altamente significativo (*F-ratio* = 4,534 con *p-value* = 0,000). Dall'osservazione della figura 3.32 sembrerebbe che il *DIF* non sia uniforme; infatti le spezzate, quasi coincidenti agli estremi, tendono a distanziarsi per i valori di abilità compresi tra 0 e 1. Tale disomogeneità tra le proporzioni empiriche non è però significativa, come risulta dal *p-value* molto alto (0,997) relativo all'interazione dei due effetti .

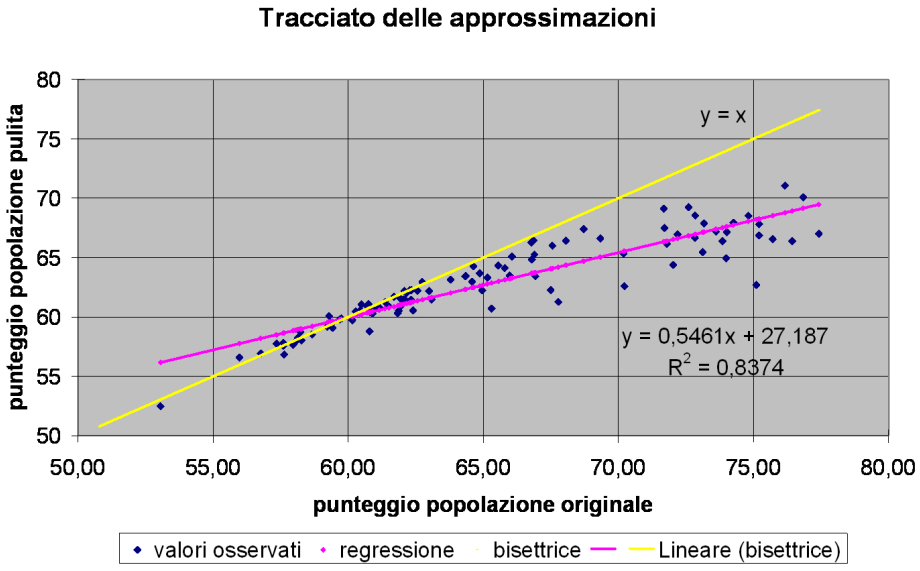


Figura 3.25: Punteggi normalizzati originali e punteggi normalizzati puliti

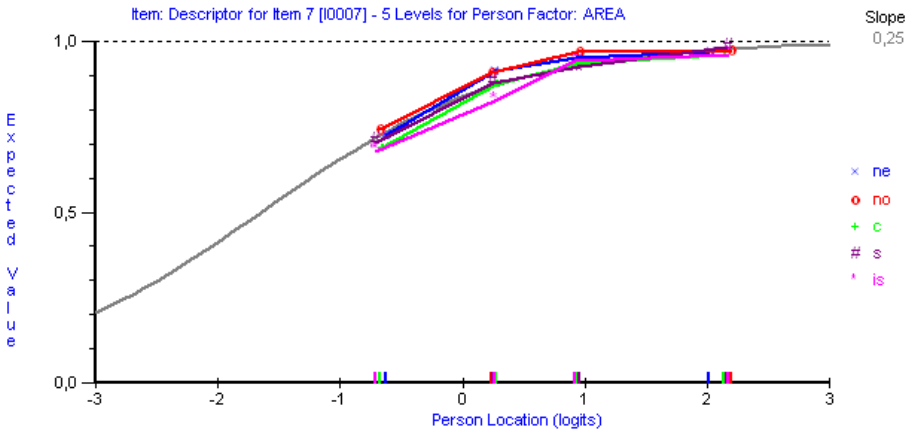


Figura 3.26: ICC dell'item 7 del test 2005/2006

SOURCE	S.S	DF	M.S	F-RATIO	Prob
Analysis of Variance for ITEM 7 [I0007:Descriptor for Item 7]					
SOURCE	S.S	DF	M.S	F-RATIO	Prob
BETWEEN	19,757	19	1,040		
ANOVA-Fit[CInt]	6,268	3	2,089	0,819524	0,483050
DIF[area]	8,895	4	2,224	0,872262	0,479808
area-by-CInt	4,594	12	0,383	0,150145	0,999650
TOTAL Item DIF	13,489	16	0,843	0,330674	0,994008
TOTAL Misfit	19,757	19	1,040	0,407861	0,988809
WITHIN	3773,299	1480	2,550		
TOTAL	3793,056	1499	2,530		

Figura 3.27: ANOVA dei residui standardizzati dell'item 7 del test 2005/2006

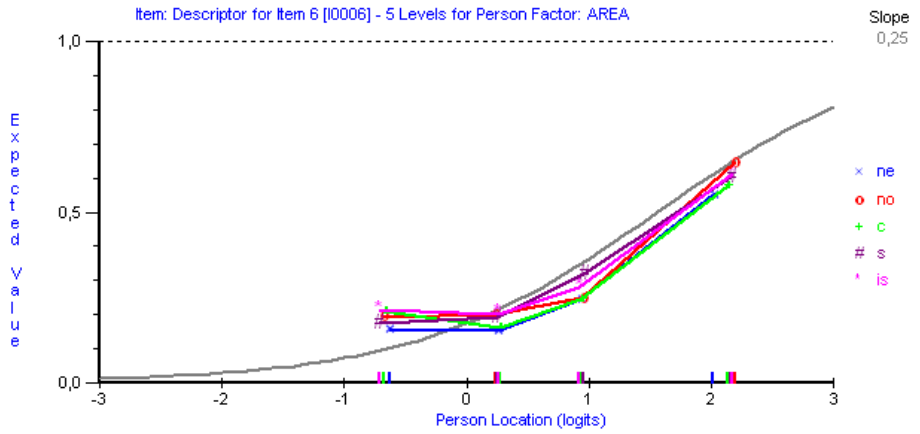


Figura 3.28: ICC dell'item 6 del test 2005/2006

SOURCE	S.S	DF	M.S	F-RATIO	Prob
Analysis of Variance for ITEM 6 [I0006:Descriptor for Item 6]					
SOURCE	S.S	DF	M.S	F-RATIO	Prob
BETWEEN	149,237	19	7,855		
ANOVA-Fit[CInt]	133,698	3	44,566	14,045430	0,000000
DIF[area]	9,225	4	2,306	0,726840	0,573596
area-by-CInt	6,314	12	0,526	0,165826	0,999413
TOTAL Item DIF	15,539	16	0,971	0,306079	0,996172
TOTAL Misfit	149,237	19	7,855	2,475451	0,000386
WITHIN	4696,031	1480	3,173		
TOTAL	4845,269	1499	3,232		

Figura 3.29: ANOVA dei residui standardizzati dell'item 6 del test 2005/2006

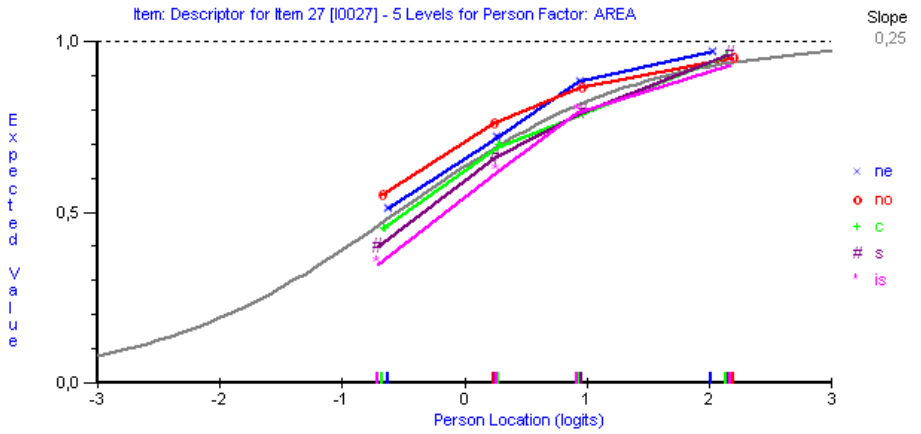


Figura 3.30: ICC dell'item 27 del test 2005/2006

Analysis of Variance for ITEM 27 [I0027:Descriptor for Item 27]

SOURCE	S.S	DF	M.S	F-RATIO	Prob
BETWEEN	54,074	19	2,846		
ANOVA-Fit[CInt]	12,593	3	4,198	1,741994	0,156465
DIF[area]	28,428	4	7,107	2,949358	0,019228
area-by-CInt	13,053	12	1,088	0,451395	0,942255
TOTAL Item DIF	41,481	16	2,593	1,075886	0,373042
TOTAL Misfit	54,074	19	2,846	1,181061	0,264839
WITHIN	3566,335	1480	2,410		
TOTAL	3620,408	1499	2,415		

Figura 3.31: ANOVA dei residui standardizzati dell'item 27 del test 2005/2006

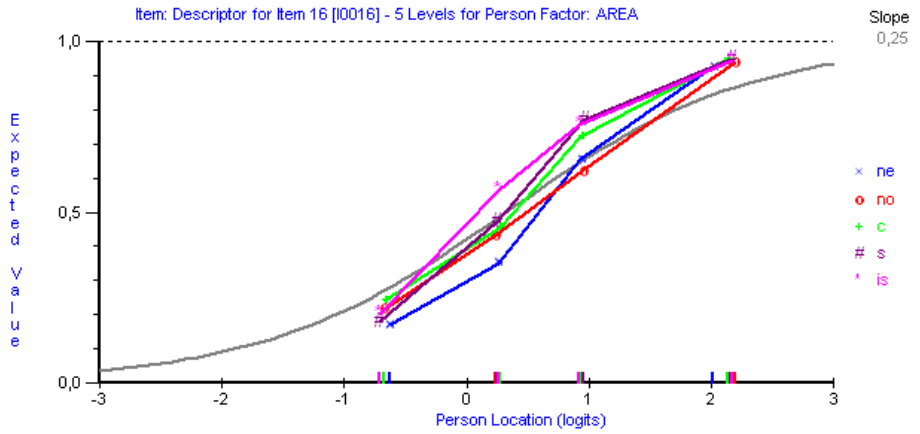


Figura 3.32: ICC dell'item 16 del test 2005/2006

SOURCE	S.S	DF	M.S	F-RATIO	Prob
Analysis of Variance for ITEM 16 [I0016:Descriptor for Item 16]					
SOURCE	S.S	DF	M.S	F-RATIO	Prob
BETWEEN	179,287	19	9,436		
ANOVA-Fit[CInt]	142,669	3	47,556	22,850370	0,000008
DIF[area]	31,054	4	7,763	3,730239	0,004984
area-by-CInt	5,564	12	0,464	0,222800	0,997412
TOTAL Item DIF	36,618	16	2,289	1,099660	0,349456
TOTAL Misfit	179,287	19	9,436	4,533983	0,000000
WITHIN	3080,186	1480	2,081		
TOTAL	3259,473	1499	2,174		

Figura 3.33: ANOVA dei residui standardizzati dell'item 16 del test 2005/2006

3.2.3 La verifica dell'efficacia della pulizia dei dati

La figura 3.25 rappresenta il grafico a dispersione dei punteggi normalizzati della popolazione originale con i punteggi normalizzati della popolazione pulita ($I_1 = 0 \ldots I_5 = 0$). Per verificare la bontà della procedura di pulizia dei database si sono innanzitutto confrontate le distribuzioni dei *total score* nelle cinque macro aree (fig. 3.34 e 3.35) prima e dopo la pulizia. È evidente che l'eliminazione dei casi di *cheating* ha condotto a distribuzioni molto più simmetriche e simili tra loro rispetto a quelle originali ma ciò non significa, di per sé, che l'affidabilità dei dati è migliorata. Per questo motivo si è voluto trovare un criterio che confermasse l'efficacia del metodo di pulizia confrontando i risultati ottenuti con i dati di una fonte informativa ufficiale, esterna al Sistema Nazionale di Valutazione. I dati più autorevoli a disposizione sono sembrati quelli delle indagini internazionali TIMSS (*Trends in International Mathematics and Science Study*), condotte dalla IEA (*International Association for the Evaluation of Educational Achievement*) di cui l'INVALSI è il referente per l'Italia che, con cadenza quadriennale, monitora gli apprendimenti della matematica e delle scienze di circa 60 paesi.

Poiché gli anni delle prove SNV oggetto di studio non coincidevano con quelli delle prove internazionali si sono utilizzati i dati TIMSS più vicini in termini temporali (cioè quelli del 2003 e del 2007), si sono individuate le scuole che hanno partecipato sia alle indagini TIMSS che alle prove del SNV del 2005/2006 e si sono resi confrontabili i punteggi delle due prove applicando agli *score* il seguente indicatore di normalizzazione:

$$min - max = \frac{x - min(x)}{max(x) - min(x)}.$$

Dopodichè è stato fatto il confronto, per ogni scuola, tra la mediana dei punteggi del TIMSS G4 (IV elementare) di matematica, prima con la mediana dei punteggi originali del SNV (senza che alcun indicatore venisse applicato) e, successivamente, con la mediana dei punteggi del SNV puliti dal *cheating* ($I1 = 0 \ldots I5 = 0$), determinando i coefficienti di regressione, l'R^2 e i coefficienti di correlazione, complessivi e suddivisi in base al genere degli scolari.

Sono state trovate 152 scuole che hanno partecipato al TIMSS 2003 e al SNV 2005/2006 (di queste, 8 sono state completamente eliminate in seguito all'applicazione degli indicatori di *cheating*) e 161 scuole che hanno partecipato al TIMSS 2007 e al SNV 2005/2006 (di queste, 15 sono state completamente eliminate in seguito all'applicazione degli indicatori di *cheating*). Le figure 3.36 e 3.37 riportano i grafici a dispersione dei dati TIMSS 2003 sui dati SNV 2005/2006, rispettivamente, prima e dopo la pulizia. Nel primo grafico la retta di regressione è perfettamente orizzontale e l'R^2 prossimo allo zero; nel secondo si percepisce un leggero miglioramento: il coefficiente angolare della retta diventa positivo e l'R^2 aumenta. L'associazione tra le due serie di dati rimane comunque molto bassa. Un discorso analogo vale se si confrontano i dati del SNV con i dati TIMSS 2007. Il grafico 3.38 indica una correlazione lineare di partenza un po' più alta rispetto a quella relativa ai dati del TIMSS del 2003, correlazione che aumenta ancora un po' dopo la pulizia dei dati, come attesta la retta di regressione della figura 3.39 in cui l'R^2 vale 0,0841. Anche in questo caso l'associazione finale rimane molto bassa. La tabella 3.5 rappresenta i coefficienti di correlazione tra le mediane delle due prove, suddivisi per genere; nella prima colonna i coefficienti sono stati calcolati sui database originari e nella seconda colonna sui database puliti. I valori molto piccoli confermano quanto detto sopra.

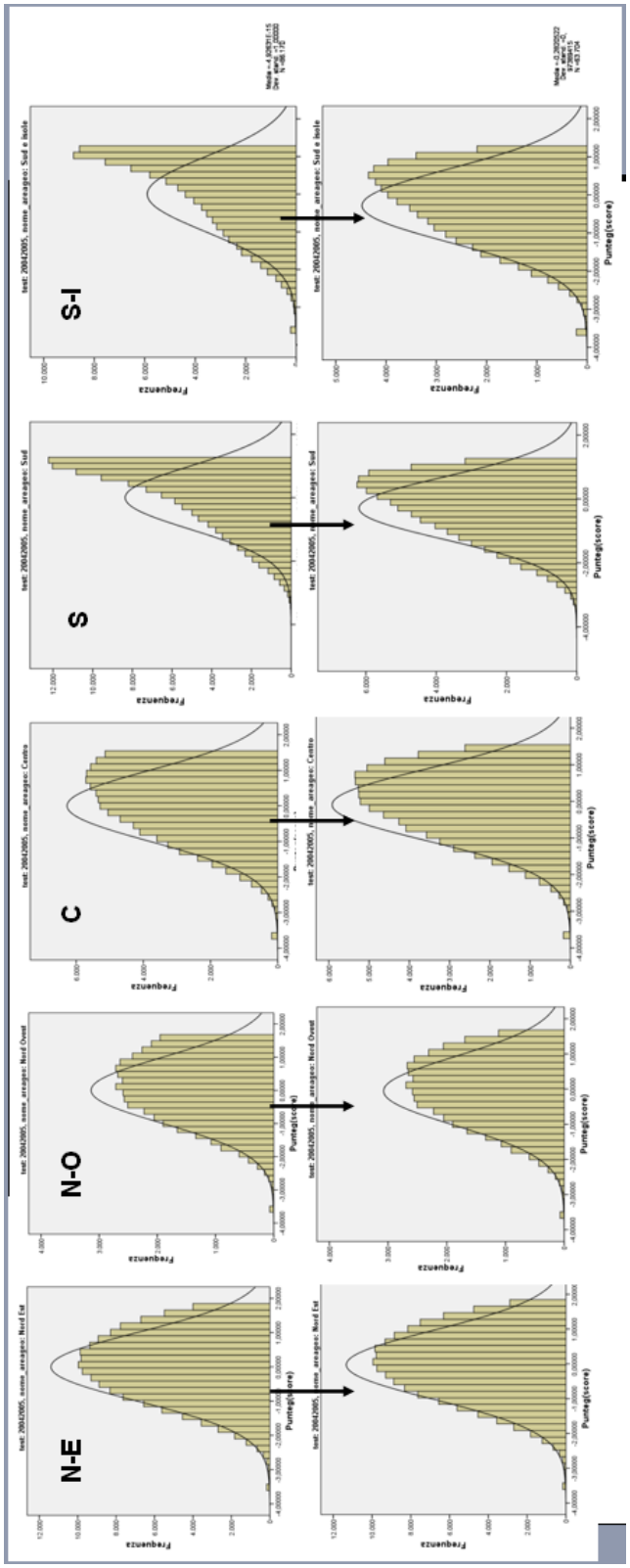


Figura 3.34: Distribuzione degli zscores del 2004/2005 prima e dopo la pulizia dei dati

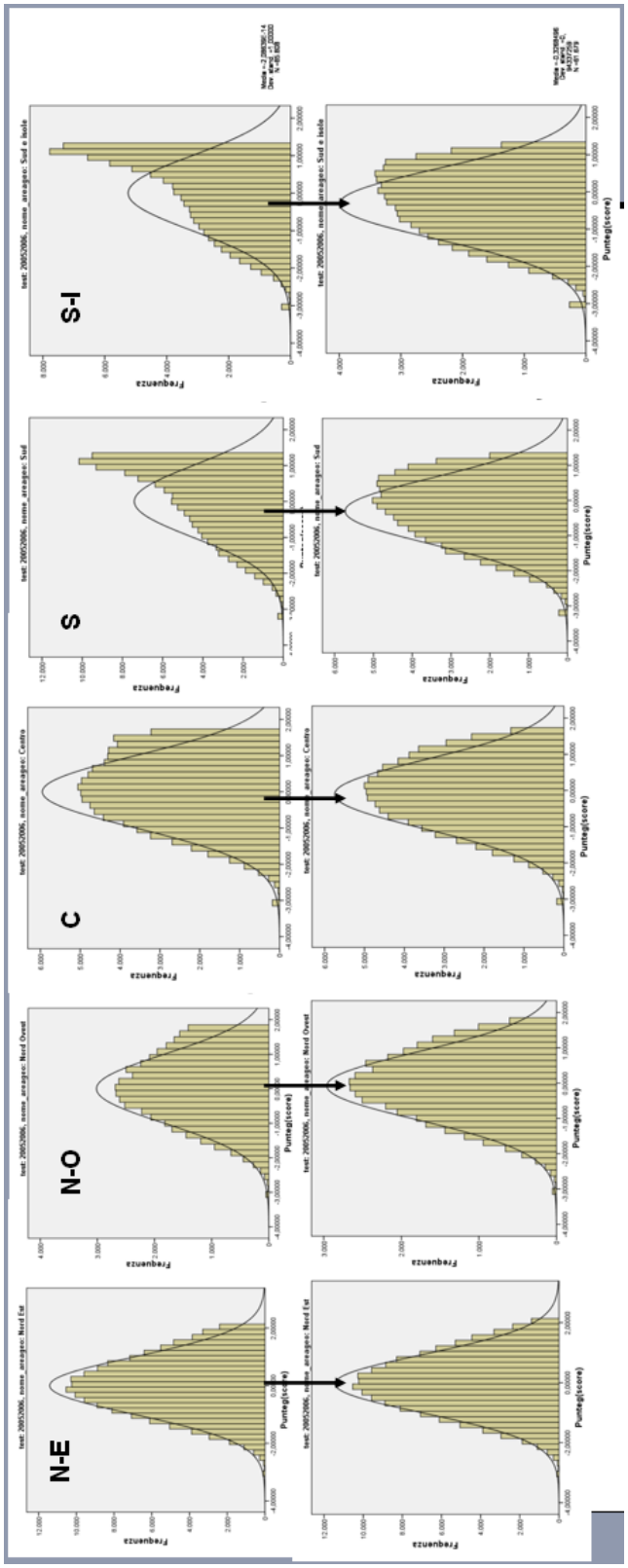


Figura 3.35: Distribuzione degli zscores del 2005/2006 prima e dopo la pulizia dei dati

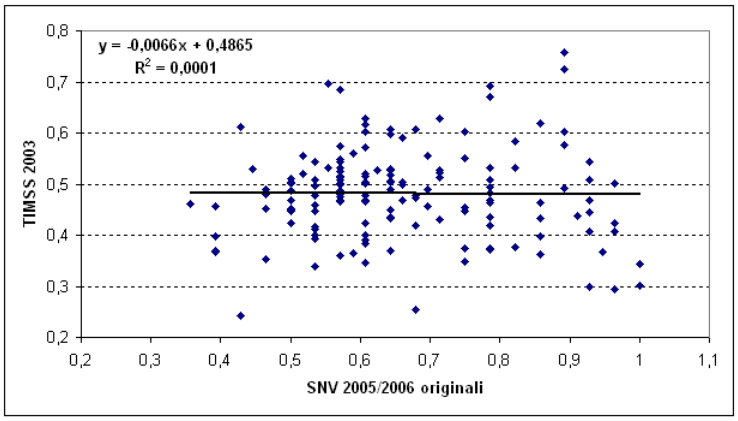


Figura 3.36: Confronto tra i dati originali del SNV 2005/2006 e i risultati TIMSS 2003

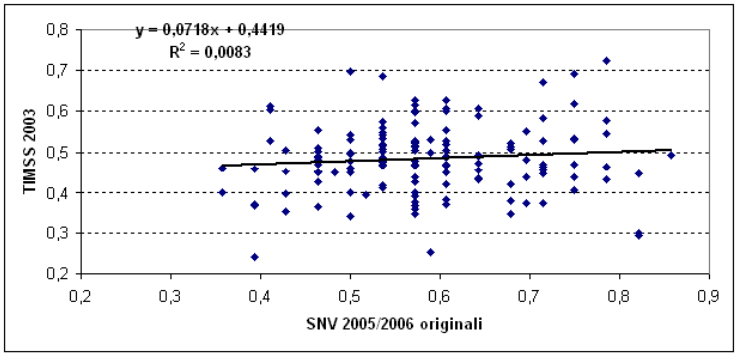


Figura 3.37: Confronto tra i dati puliti del SNV 2005/2006 e i risultati TIMSS 2003

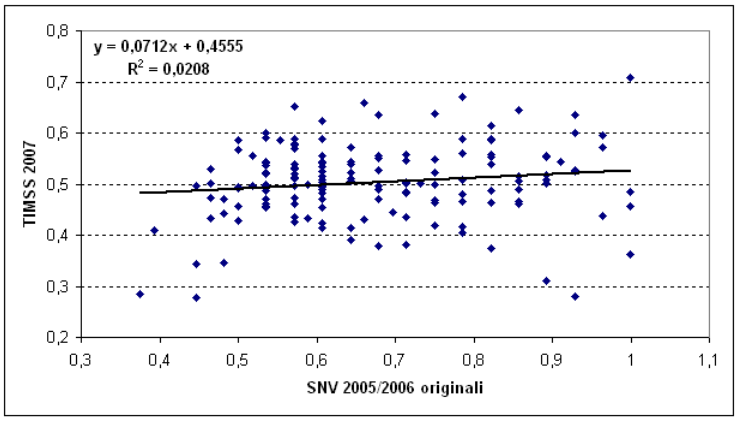


Figura 3.38: Confronto tra i dati originali del SNV 2005/2006 e i risultati TIMSS 2007

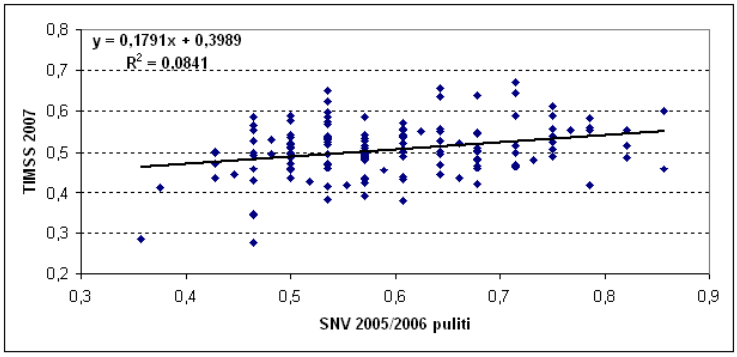


Figura 3.39: Confronto tra i dati puliti del SNV 2005/2006 e i risultati TIMSS 2007

TIMSS 2003 vs SNV 0506 originale			TIMSS 2003 vs SNV 0506 pulito		
maschi			maschi		
	TIMSS	SNV		TIMSS	SNV
TIMSS	1		TIMSS	1	
SNV	-0,04	1	SNV	0,06	1

femmine			femmine		
	TIMSS	SNV		TIMSS	SNV
TIMSS	1		TIMSS	1	
SNV	-0,03	1	SNV	0,04	1

maschi + femmine			maschi + femmine		
	TIMSS	SNV		TIMSS	SNV
TIMSS	1		TIMSS	1	
SNV	-0,01	1	SNV	0,09	1

TIMSS 2007 vs SNV 0506 originale			TIMSS 2007 vs SNV 0506 pulito		
maschi			maschi		
	TIMSS	SNV		TIMSS	SNV
TIMSS	1		TIMSS	1	
SNV	0,13	1	SNV	0,27	1

femmine			femmine		
	TIMSS	SNV		TIMSS	SNV
TIMSS	1		TIMSS	1	
SNV	0,14	1	SNV	0,22	1

maschi + femmine			maschi + femmine		
	TIMSS	SNV		TIMSS	SNV
TIMSS	1		TIMSS	1	
SNV	0,14	1	SNV	0,29	1

Tabella 3.5: Correlazione tra gli score del SNV 2005/2006 e i dati TIMSS; maschi, femmine, maschi+femmine

Capitolo 4

La costruzione del test di *link* e della scala di misura

4.1 La predisposizione del test

4.1.1 La definizione

Per la costruzione del test di *link* ci si è avvalsi della gentile e preziosa collaborazione della professoressa Maria Pia Perelli, docente di Matematica alla facoltà di Scienze della Formazione dell'Università degli Studi di Trieste, esperta di didattica e di valutazione degli apprendimenti della matematica e consulente dell'INVALSI. Si è convenuto che il test di *link* dovesse valutare le conoscenze e le competenze di un insieme di concetti fondamentali della matematica e la capacità di comprendere il linguaggio logico, preciso e coerente, con cui i quesiti matematici vengono posti, sulla base di quelle che sono state le indicazioni dei gruppi di lavoro che hanno preparato le prove di valutazione dell'INVALSI negli ultimi anni.

L'idea portante dei gruppi di lavoro è stata di trovare strumenti per individuare, nel percorso formativo di ogni livello scolastico, delle conoscenze elementari che fossero espressione della matematica come fattore di crescita per la persona, strumento di conoscenza della realtà nonché linguaggio preciso, univoco, obiettivo e utile, se non indispensabile, per descrivere tale realtà, evitando di eccedere in astrazioni e formalismi. Si è voluto, cioè, sottolineare

l'importanza di una concezione della matematica che fosse "1) indipendente dagli stereotipi suggeriti sia dalla evidenza intuitiva, sia dalle immagini mentali memorizzate in modo acritico, sia dagli automatismi dell' addestramento algoritmico e 2) attuata in contesti critici di razionalizzazione della realtà: ciò per avere indicazioni sia sul livello di un'appropriazione personale critica e interiorizzata della conoscenza stessa sia sull'abilità nell'uso di alcuni strumenti (=algoritmi) matematici elementari, ma cruciali, nel ruolo di descrizione e di controllo (modellizzazione) della realtà" (Mario Marchi, 2004). Pertanto rispondere alle domande inserite nel test non ha richiesto solo di eseguire calcoli o applicare formule più o meno note ma anche di fare appello all'intuizione, al ragionamento, alle conoscenze e alle relative abilità acquisite. Inoltre, anche quando la domanda si riferiva all'esecuzione di procedimenti algoritmici, la difficoltà non è mai consistita nella lunghezza o nella complicazione dei calcoli, bensì nella necessità di possedere con chiarezza e sicurezza i concetti implicati.

Visto che sia il questionario di rilevazione degli apprendimenti dell'anno scolastico 2004/2005 sia quello dell'anno 2005/2006 sono risultati mediamente troppo facili (vedi capitolo precedente), nel test di *link*, a fianco alle domande di collegamento prese dalle due prove summenzionate, sono stati inseriti dei quesiti presi dalla prova TIMSS del 2007 relativa al quarto anno delle scuole primarie, generalmente più difficili dei quesiti nazionali perché colgono competenze che non vengono insegnate nel nostro sistema scolastico, nonché quesiti desunti tra gli item scartati (nonostante fossero stati valutati dagli esperti come "buoni" item) dalle ultime prove del Servizio Nazionale di Valutazione somministrate per monitorare gli apprendimenti della V classe della scuola primaria; sono stati, infine, utilizzati cinque item presenti in vecchie prove di valutazione della matematica nella I classe della scuola secondaria di primo grado. Si è ritenuto, infatti, che tali domande, nonostante fossero concettualmente più sottili, e quindi più difficili, delle altre, potessero essere svolte (e così è stato) da buona parte dei ragazzi dotati di conoscenze e competenze matematiche superiori alla media.
Per mantenere un'uniformità tra le domande e rendere più agevole l'inserimento e la pulizia dei dati, tutti i 32 quesiti da cui è composto il test di *link* sono stati scelti a risposta chiusa, con una sola risposta corretta e tre distrattori.

Allo scopo di testare il questionario, stimare la difficoltà delle domande e, successivamente, valutare la bontà del *link* tra i questionari del SNV sono stati presi 422 alunni di 9 scuole, suddivisi come riportato nella tabella seguente:

Scuola	
Circolo Didattico di Latisana (UD)	109
Istituto Comprensivo di Premariacco (UD)	88
Scuola Primaria Parificata Collegio della Provvidenza (UD)	15
Scuola Elementare Collegio delle Dimesse (UD)	25
Scuola Primaria Bearzi (UD)	19
Scuola Primaria Bertoni (UD)	23
Scuola Primaria Morpurgo (TS)	114
Scuola di Venezia (VE)	29
	422

4.1.2 I temi e i contenuti delle domande

Passando a illustrare le conoscenze e abilità valutate con il questionario, in generale, le 32 domande hanno cercato di sondare la capacità di:

- interpretare correttamente un testo, anche in presenza di connettivi logici;
- eseguire calcoli (non eccessivamente complicati) e riconoscere le operazioni e i procedimenti adatti a risolvere problemi specifici;
- saper utilizzare e interpretare un formalismo simbolico in un contesto assegnato;
- fare ed esprimere deduzioni e riconoscere i collegamenti logici;
- "leggere" diverse forme di rappresentazione e interpretare in modo corretto le informazioni in esse contenute.

I temi e i contenuti proposti sono stati quelli dei questionari di valutazione somministrati nel 2004/2005 e nel 2005/2006 e precisamente i seguenti:

Tema 1: **I Numeri**: conoscere, confrontare e ordinare numeri naturali e decimali, operare con i numeri naturali e decimali, riconoscendo le proprietà

delle operazioni, saper stimare il risultato di un'operazione, saper utilizzare la frazione come operatore;

Tema 2: **Geometria**: conoscere e denominare alcune importanti figure geometriche piane, conoscere le principali proprietà delle figure geometriche piane, individuare simmetrie in oggetti e figure, saper determinare l'area e il perimetro di semplici figure;

Tema 3: **Misure**: saper riconoscere l'ordine di grandezza di una misura, conoscere il sistema metrico decimale e saper effettuare semplici conversioni tra un'unità di misura e un'altra;

Tema 4: **Dati e previsioni**: saper leggere e ricavare informazioni da un diagramma a barre e da un areogramma circolare;

Tema 5: **Introduzione al pensiero razionale**: utilizzare in modo corretto alcuni termini della matematica, saper classificare figure, oggetti, numeri, saper individuare le informazioni necessarie e saper organizzare un percorso adeguato per risolvere un problema.

Ogni domanda ha riguardato uno o più temi e contenuti e per ogni tema si è cercato di porre almeno un quesito facile, uno di media difficoltà e uno difficile. Ecco alcuni esempi di domanda.

Esempio

Domanda 21. Quale dei seguenti numeri corrisponde a "tre decine di migliaia, tre centinaia e tre decimi"?
A. 3 333,3
B. 300 300,3
C. 30 300,3
D. 30 300,03

Questo item voleva testare la capacità di conoscere non solo le migliaia, le centinaia, le unità e i numeri decimali ma anche la lettura e la scrittura degli stessi.

Esempio

Domanda 12. Aldo misura la lunghezza di una lavagna con un righello

di 30 cm. La lavagna misura 6 cm in meno rispetto a 9 volte la lunghezza del righello. Qual è la lunghezza della lavagna?
A. 264 cm
B. 270 cm
C. 276 cm
D. 279 cm

Il problema aveva lo scopo di far emergere la capacità di interpretare un testo comprendendo anche tutti i termini specifici (quali "9 volte la ...") e di eseguire le operazioni corrette, nel modo e nell'ordine giusto.

Esempio

Domanda 3. Quale dei seguenti prodotti NON dà come risultato 90?
A. 45 x 2 x 1
B. 9 x 2 x 5
C. 3 x 5 x 6
D. 8 x 3 x 5

Questo item richiedeva la capacità di eseguire le moltiplicazioni nonché quella di interpretare correttamente la domanda in presenza di un operatore logico (NON).

Si è anche tenuto conto del fatto che gli esiti delle prove sono sempre stati legati ai "prodotti" e non ai "processi" di apprendimento, come si vede dall'esempio che segue.

Esempio

Domanda 16. 23 x 9 dà un risultato minore di 23 x 10; di quanto minore?
A. 1
B. 9
C. 10
D. 23

Ci sono almeno due modi per rispondere a questa domanda: il primo, più elegante, e che richiede una buona conoscenza delle proprietà delle operazioni,

No item	Tema	Contenuto	Difficoltà	Provenienza
1	Numeri	Operazioni	F	D28 (IV e 2005/06)
2	Dati e previsioni	Lettura di grafico a barre	M	D01 (IV e 2004/05)
3	Numeri	Operazioni	F/M (per il NON nella domanda)	D24 (IV e 2005/06)
4	Geometria	Riconoscimento figure	D (per la presenza della figura 2)	D15 (IV e 2005/06)
5	Geometria e logica	Riconoscimento figure	M	D27 (IV e 2004/05)
6	Numeri	Calcolo non decimale	M	D03 (I m 2005/06)
7	Geometria	Riconoscimento perimetri e aree	M/D	D03 (I m 2004/05)
8	Dati e previsioni	Lettura di grafico	F/M	D27 (I m 2005/06)
9	Numeri	Operazioni e valore posizionale delle cifre	M/D	TIMSS
10	Numeri	Riconoscimento frazioni	F	TIMSS
11	Geometria	Calcolo area	F	TIMSS
12	Misura	Confronto tra diverse unità di misura	D	TIMSS
13	Numeri	Proprietà operazioni	F	Scarto V
14	Geometria	Denominazione figure	F	D09 (IV e 2005/06)
15	Numeri	Proprietà dei numeri	M/D	Scarto V
16	Numeri	Proprietà dei numeri	M	D13 (IV e 2005/06)
17	Numeri	Problema con decimali	M	D22 (IV e 2004/05)
18	Numeri	Lettura e scrittura	M/D	D21 (IV e 2005/06)
19	Numeri	Problema con operazioni	F	TIMSS
20	Numeri	Sequenze	D	TIMSS
21	Numeri	Scrittura metrica decimale	M	Scarto V
22	Numeri	Rappresentazione simbolica	D	TIMSS
23	Geometria/logica	Proprietà figure	D	Scarto V
24	Dati e previsioni	Lettura tabella	D (per il “più”)	D13 (I m 2005/06)
25	Misure	Equivalenze e confronti	F	D11 (IV e 2005/06)
26	Numeri	Confronto tra numeri e stima	M/D	TIMSS
27	Numeri	Uso delle frazioni	F/M (“quarto” scritto in lettere)	D25 (IV e 2004/05)
28	Dati e previsioni	Lettura grafici e confronto	D (per il calcolo)	TIMSS
29	Numeri/misure	Ordine di grandezza	M	D30 (I m 2005/06)
30	Misura	Confronto tra diverse unità di misurazione	D	Scarto V
31	Geometria	Riconoscimento simmetrie	M (per il rettangolino con la diagonale)	Scarto V
32	Geometria	Riconoscimento parallelismo	F	D23 (IV e 2004/05)

Tabella 4.1: I temi e i contenuti degli item del test di link

consiste nel riconoscere che 23 x 10 = 23 x (9 + 1) = 23 x 9 + 23 x 1 = 23 x 9 + 23; il secondo consiste nel calcolare i risultati di 23 x 9 e di 23 x 10 e poi determinare qual è la differenza tra i due.

Nella tabella 4.1 per ognuna delle 32 domande è riportato il tema (o i temi) e il contenuto (o i contenuti) cui si riferisce, nonché il suo livello di difficoltà presunta (facile, medio, difficile) e la sua provenienza (SNV 2004/2005 o 2005/2006, scarti SNV V elementare, vecchi SNV I media, TIMSS).

4.2 L'analisi del test

Ci si soffermerà ora a valutare l'adeguatezza del questionario somministrato per misurare le abilità degli scolari e per fungere da test di *link*. Per realizzare tali obiettivi si è fatto ampio uso delle analisi contenute in RUMM2020 e in Winsteps. Le prime analisi sono state svolte considerando tutti gli item e tutti gli scolari; quelle finali si sono focalizzate solo sui casi che si adattavano al RM.

La figura 4.1 mostra la finestra che compare non appena RUMM2020 ha eseguito la stima dei parametri ed è pronto a elaborare le analisi e a mostrarne i risultati. Le sezioni più interessanti nella prospettiva qui assunta sono quelle relative all'adattamento globale (*Test-of-Fit Details*), all'analisi delle *ICC* (*Item Characteristics*) e alla rappresentazione grafica delle stime dei parametri (*Further Outputs*).

4.2.1 L'affidabilità delle osservazioni

Dalla sezione del *Test-of-Fit details* si accede ai risultati delle principali statistiche di adattamento dei dati al modello, sia a livello globale (*Summary Statistics*), sia a livello dei singoli item (*Individual Item Fit*), sia a livello dei singoli soggetti (*Individual Person Fit*).

La figura 4.2 mostra la finestra dell'output delle statistiche globali del test somministrato nell'anno scolastico 2008/2009 (test di *link*) agli allievi della classe IV della scuola primaria. Nella parte alta della finestra (*Item-Person Interaction*) sono contenute le statistiche di sintesi delle stime dei parametri e dei loro residui. Di *default* la media delle difficoltà degli item è posta ar-

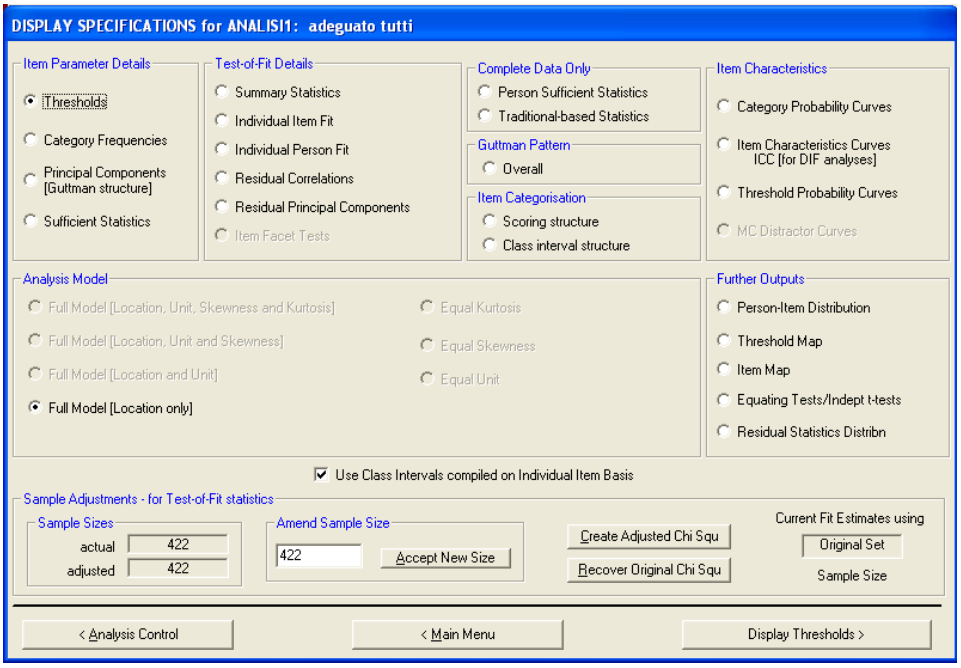


Figura 4.1: La finestra Display Specifications di RUMM2020

bitrariamente a 0 per fissare un'origine della scala di misurazione (è possibile modificare tale valore o fissare l'origine sulla media delle abilità stimate, ma non se ne è mai fatto uso nel presente lavoro). La deviazione standard delle difficoltà stimate è pari a 1,173; tale valore costituisce l'unità di misura della scala, determinata dai dati utilizzati per la sua costruzione. La media delle abilità stimate è 0,307, il che significa che le stime delle capacità degli scolari è di circa 0,3 logit superiore a quella degli item (il test è risultato leggermente troppo facile per gli individui monitorati nonostante siano state inserite domande relative a livelli di competenza superiori a quelli di IV elementare; ad ogni modo il test di link è stato calibrato meglio delle prove SNV 2004/2005 e 2005/2006), la loro *sd* è 0,979.

I residui forniscono indicazioni sul livello di adattamento dei dati e sono costruiti in maniera tale da approssimare una distribuzione normale standardizzata. Nel caso in cui i dati siano conformi al modello, gli indicatori sintetici della loro distribuzione devono essere vicini a quella di una normale di media nulla e varianza unitaria. Sia i residui degli item che quelli delle persone si

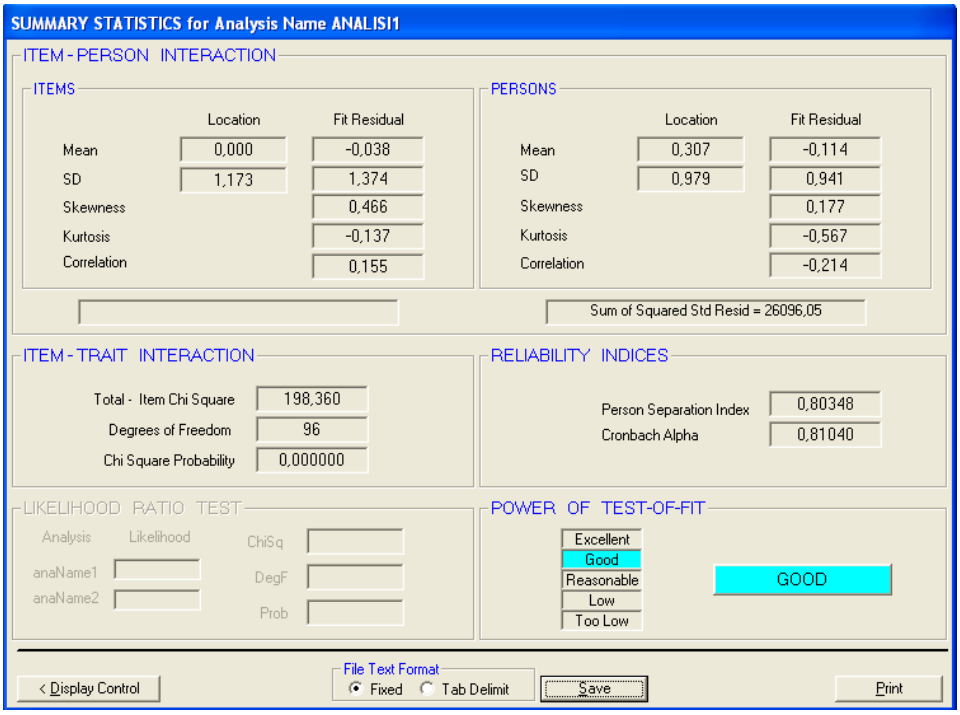


Figura 4.2: Summary statistics

scostano da una distribuzione normale. Quelli degli item hanno una media prossima allo zero, una SD pari a 1,374, una *Skewness* positiva uguale a 0,466 e una curtosi negativa uguale a -0,137: la distribuzione è asimmetrica positiva e leggermente platicurtica. Quelli delle persone hanno una media leggermente negativa (-0,114), una *sd* prossima all'unità (0,941), una *Skewness* leggermente positiva (0,177) e una curtosi negativa (-0,567): la distribuzione è leggermente asimmetrica positiva e platicurtica.

La correlazione tra i residui, infine, serve a verificare se esiste un legame tra le risposte degli item: un alto valore positivo indica che c'è una forte correlazione tra le risposte date dagli studenti ai vari item e che quindi si potrebbe supporre che tra le domande c'è una dipendenza dovuta a una qualche forma di ridondanza. Il valore osservato è leggermente positivo (0,155) e non fa supporre nessuna anomalia degli item in questo senso. La correlazione dei residui tra gli studenti è leggermente negativa (-0,214) e anche in questo caso va considerata come fisiologica. La finestra riporta anche due indici che esprimono il

grado di affidabilità con cui il test riesce a separare i soggetti: il *Person Separation Index* (Andrich, 1982a) e il *Cronbach's Alpha*; entrambi rappresentano la proporzione di variabilità osservata dovuta alla variabilità effettiva (variabilità fra le abilità latenti stimate). In caso di dati completi (come in questo caso) tutti e due gli indici vengono calcolati e forniscono approssimativamente lo stesso risultato, in caso di presenza di valori mancanti solamente il *Person Separation Index* (*PSI*) può essere calcolato. Sull'indice *Cronbach's Alpha* si è avuto modo di parlare a lungo nel paragrafo 1.1.2. Richiamiamo la formula per calcolarlo.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\widehat{\sigma}_i^2}{\widehat{\sigma}_X^2}\right) = \frac{k}{k-1}\left(\frac{\widehat{\sigma}_X^2 - \sum_{i=1}^{k}\widehat{\sigma}_i^2}{\widehat{\sigma}_X^2}\right)$$

dove

- k è il numero degli item del test;
- $\widehat{\sigma}_X^2 = \frac{1}{n-1}\sum_{\nu=1}^{N}\left(X_\nu - \overline{X}\right)^2$ è la stima della varianza totale del test;
- $\widehat{\sigma}_i^2 = \frac{1}{n-1}\sum_{\nu=1}^{N}\left(X_{\nu i} - \overline{X_i}\right)^2$ è la stima della varianza di ciascun item.

Nel RM l'indice di affidabilità è determinato partendo dall'ipotesi che, per ogni soggetto ν, l'abilità osservata delle persone sia la somma di due componenti: l'abilità effettiva (θ_ν), incognita, e l'errore di misurazione ϵ_ν, cioè:

$$\widehat{\theta_\nu} = \theta_\nu + \epsilon_\nu. \tag{4.1}$$

con $E(\widehat{\theta_\nu}) = \theta_\nu$ e $Var(\widehat{\theta_\nu}) = \sigma_\epsilon^2$

Se si assume che la popolazione di studio abbia media μ e varianza σ_θ^2 si può scrivere:

$$\theta_\nu = \mu + \phi_\nu \tag{4.2}$$

con ϕ_ν distanza tra il soggetto ν e la media μ, da cui segue che $E(\theta) = \mu$ e $Var(\theta) = Var(\phi) = \sigma_\theta^2$. La stima dell'abilità del soggetto ν diventa:

$$\widehat{\theta_\nu} = \mu + \phi_\nu + \epsilon_\nu. \tag{4.3}$$

con $E(\widehat{\theta}) = \mu$ e $\sigma^2_{\widehat{\theta}} = Var(\widehat{\theta}) = \sigma^2_\theta + \sigma^2_\epsilon$. L'indice di separazione è dato dalla:

$$r_\theta = \frac{\sigma^2_\theta}{\sigma^2_{\widehat{\theta}}} = \frac{\sigma^2_{\widehat{\theta}} - \sigma^2_\epsilon}{\sigma^2_{\widehat{\theta}}} = 1 - \frac{\sigma^2_\epsilon}{\sigma^2_{\widehat{\theta}}} \tag{4.4}$$

dove

- $\sigma^2_{\widehat{\theta}} = \frac{1}{n-1}\sum_{\nu=1}^{n}(\widehat{\theta}_\nu - \overline{\widehat{\theta}}_\nu)^2$ è la stima della varianza delle abilità stimate e
- $\widehat{\sigma}^2_\epsilon = \frac{1}{n}\sum_{\nu=1}^{n}\widehat{\sigma}^2_{\epsilon_\nu}$ è la stima della varianza dell'errore, che si utilizza in pratica nella 4.4.

Il *PSI*, a differenza del *Cronbach's Alpha*, non considera il numero degli item nel test. Come l'indice α, però, fornisce una misura della coerenza interna dei risultati; indica, cioè, sulla base dei risultati osservati, in che misura gli item sono in grado di discriminare tra gli studenti. Né l'α di Cronbach né il *PSI* forniscono di per sé misure circa la bontà (in termini di "unidimensionalità" e di "indipendenza locale") degli item; basti pensare che per aumentare il valore di entrambi gli indici è sufficiente dare due volte gli stessi item (vedi la *Spearman-Brown prophecy formula*), violando manifestamente il presupposto dell'indipendenza locale.

Il valore dell'indice di separazione per le osservazioni del test di *link* è 0,803 (di poco inferiore al *Cronbach's Alpha* che vale 0,810), da cui si può dedurre che gli item del test riescono a discriminare bene tra le diverse abilità degli studenti del campione. La verifica della validità dello strumento (il test) richiede altre analisi più specifiche. La casella del *Power of Test of Fit* non è nient'altro che un giudizio sul livello di fiducia che si può dare al test di adattamento dei dati al modello basato esclusivamente sul valore del *PSI*. Maggiore è il valore del *PSI* maggior fiducia si può riporre nei risultati del test di adattamento.

4.2.2 Valutazione generale del test

La prima verifica è stata osservare come gli item e le persone si distribuiscono lungo il *continuum*. La figura 4.3 rappresenta la finestra di RUMM2020 contenente la scala di misura su cui sono poste le due distribuzioni: gli item

nella parte inferiore e gli studenti nella parte superiore. A sinistra sono collocate le domande più facili; man mano che ci si sposta verso destra la difficoltà aumenta fino a raggiungere gli item più impegnativi. Analogamente, nella parte più alta del grafico, a sinistra si trovano gli alunni con più scarse competenze matematiche e a destra quelli con competenze maggiori.
Sul grafico sono riportati la numerosità del campione (422), la media (0,307) e la *standard deviation* (0,979) delle stime delle abilità dei soggetti. Come già detto, la media superiore allo zero indica che il test è risultato relativamente facile per il campione a cui è stato somministrato; in effetti dalle due distribuzioni si evince che tutti gli alunni hanno almeno un item sul livello della scala in cui si trovano, fatta eccezione per i pochi alunni più bravi che invece risultano "scoperti".

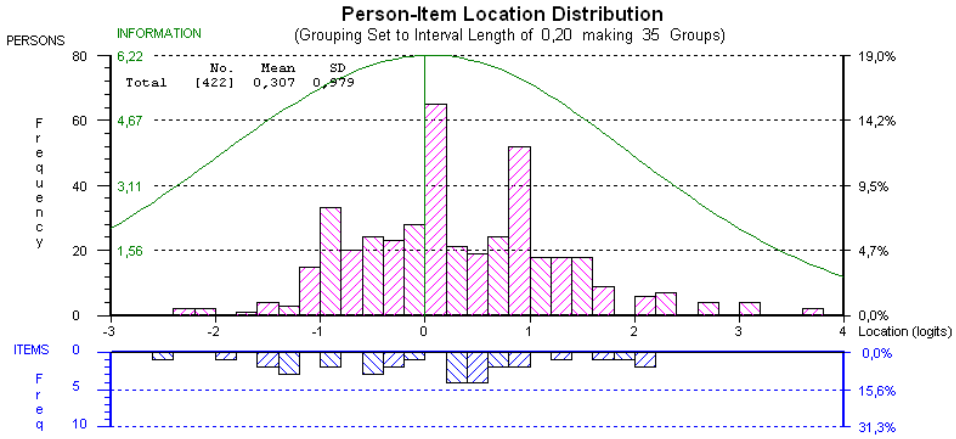


Figura 4.3: Item Person Distribution

La curva disegnata sopra la distribuzione degli item rappresenta la curva di informazione del test. Accanto al concetto di affidabilità (generalmente espressa dal rapporto tra la varianza del *true score*, o dell'abilità reale, e la varianza dell'*observed score*, o dell'abilità stimata) c'è il concetto di precisione che si riferisce all'accuratezza della stima . Nell'*IRT* la precisione delle stime non è costante su tutto il *range* dei punteggi ma diminuisce man mano che ci si avvicina ai valori estremi della scala dove l'errore di misurazione è maggiore. Il concetto di informazione in ambito statistico, introdotto per la prima volta da

Ronald Fisher, è collegato a quello di errore standard della misurazione. Infatti un indicatore della precisione della stima di un parametro è rappresentato dalla variabilità della stima attorno al suo valore "vero". Si può quindi pensare che tanto maggiore è la variabilità tanto minore è l'informazione della stima; ciò può essere espresso attraverso la formula $I = 1/\sigma^2$

Nell'*IRT* l'interesse primario è la misurazione di caratteristiche latenti dei soggetti; in questo caso quindi, per ogni soggetto, l'informazione è data dal reciproco della variabilità della stima della sua abilità ed essendo l'abilità una variabile continua, anche l'informazione è una variabile continua.
L'informazione, inoltre, può essere calcolata per ogni item o per l'intero test. La funzione che ne rappresenta i valori è la Funzione di Informazione (*Information Function*) che generalmente viene indicata con $I_i(\theta)$. Nel *SLM* la Funzione di Informazione è semplicemente il prodotto tra la probabilità di una risposta corretta e la probabilità di una risposta sbagliata

$$I_i(\theta) = P_i(\theta)Q_i(\theta) \tag{4.5}$$

dove

- $P_i\left(\theta\right) = \dfrac{exp(\theta - \delta_i)}{1 + exp(\theta - \delta_i)}$,
- $Q_i(\theta) = 1 - P_i(\theta)$,
- θ è il livello di abilità.

La funzione di Informazione del Test indica la massima accuratezza con cui il parametro dell'abilità di una persona può essere stimato lungo tutto il *range* delle abilità. Valendo l'ipotesi di indipendenza locale la funzione di informazione del test è data dalla somma delle funzioni di informazione dei singoli item:

$$I(\theta) = \sum_{i=1}^{k} I_i(\theta) \tag{4.6}$$

In generale l'informazione del test è maggiore dell'informazione di ogni singolo item e maggiore è il numero degli item maggiore è il livello di informazione relativo a ciascun livello di abilità. Nella figura 4.3 si vede che la funzione di

informazione del test è massima attorno allo zero, decresce man mano che ci si allontana dall'origine pur mantenendosi a livelli alti tra -2 e +2, dove ricade la maggior parte degli allievi. Lo *standard error* della stima è il reciproco della Funzione di Informazione del Test

$$SE(\theta) = \frac{1}{\sqrt{(I(\theta))}}.$$

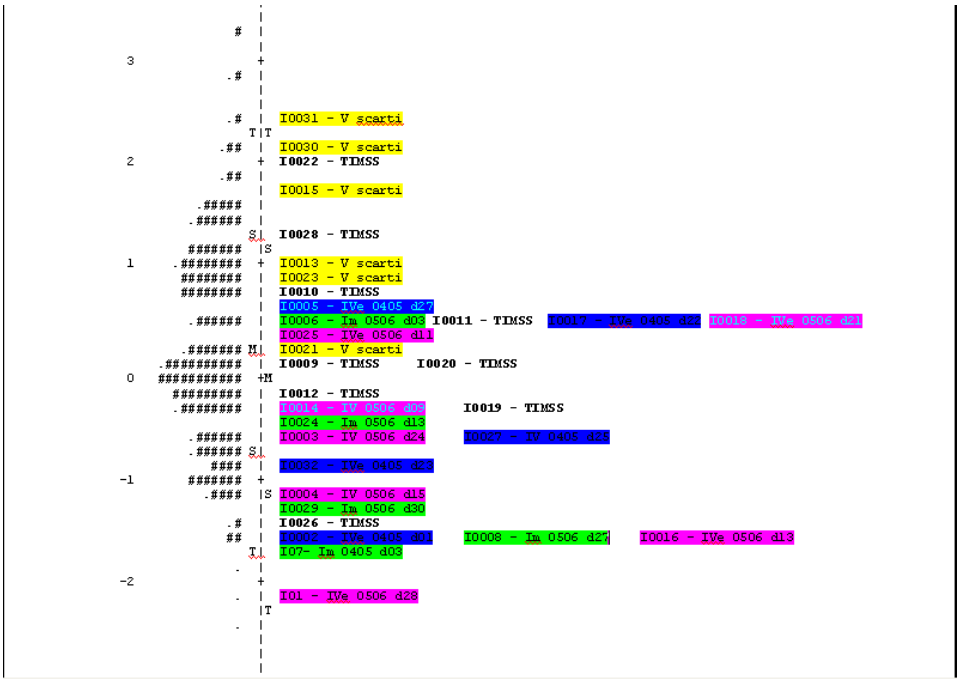


Figura 4.4: Item-Person Distribution Map del test di link

4.2.3 Le Item Characteristic Curves e il Fit Residual

In questo paragrafo si prendono in considerazione le principali analisi svolte per testare l'adeguatezza dei singoli item a misurare l'abilità latente degli studenti. In particolar modo l'attenzione si focalizza sull'analisi delle Curve Caratteristiche degli Item (*Item Characteristic Curve*, in breve *ICC*), sullo studio dei distrattori e sull'osservazione di altri indicatori statistici. In tutta l'*Item Response Theory* (a 1, a 2 e a 3 parametri) lo studio delle *ICC* è di

centrale importanza. La *ICC* di un item rappresenta la probabilità di rispondere correttamente all'item in funzione del livello di abilità stimato. Nel RM il raffronto tra la curva stimata e la curva empirica osservata fornisce la prima indicazione sulla bontà dell'item.

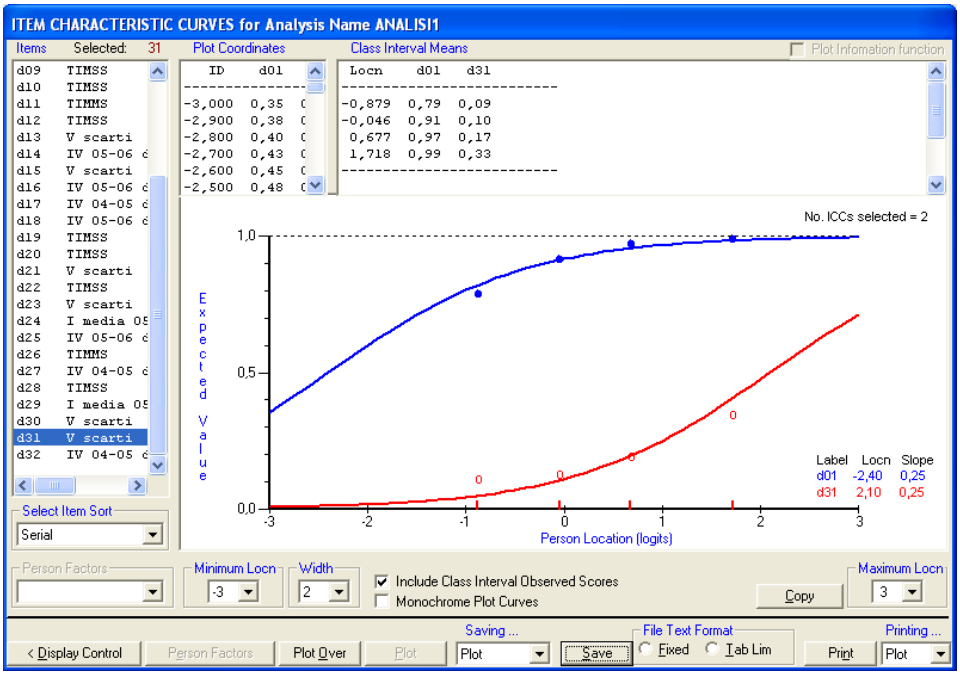


Figura 4.5: ICC degli item d01 e d31

La figura 4.5 rappresenta l'output di RUMM2020 per l'analisi delle ICC. Nell'elenco a sinistra, sotto la voce item, compare la lista delle domande del test da cui è possibile selezionare l'item o gli item da rappresentare sul grafico. Più a destra, in alto, compaiono le coordinate delle ICC (*Plot Coordinates*) e ancora più a destra le frequenze medie di risposte corrette osservate (*Class Interval Means*) in ciascuna delle quattro classi di intervallo in cui è stato suddiviso il campione. Il grafico rappresenta le ICC dell'item più facile (l'item 1) e dell'item più difficile (l'item 31) del test; le linee continue rappresentano le curve teoriche, i puntini i valori osservati delle medie delle quattro classi. Se, per esempio, si vuole sapere qual è la probabilità, per una persona di abilità stimata pari a 0 (quindi un'abilità uguale alla difficoltà media delle domande), di rispondere in maniera corretta ai due item, è sufficiente tracciare una linea

ipotetica verticale in corrispondenza dell'origine della scala e osservare per quali valori dell'asse delle ordinate tale linea interseca le due curve; nel nostro caso in prossimità di 0,9 per l'item più facile e di 0,1 per quello più difficile. La difficoltà di ogni item corrisponde, sulla scala, al livello in cui la probabilità di rispondere correttamente è pari a 0,5; per l'item 1 è -2,40 e per l'item 31 è +2,10 (come si può vedere dal riquadro in basso a destra sul grafico dove compaiono l'etichetta, la *location* e la *slope* dei due item). Conoscendo la difficoltà dell'item e il livello di abilità è possibile determinare con esattezza la probabilità di rispondere correttamente stimata dal modello: basta sostituire i due valori nella 1.34. Così per $\theta = 0$ e $\delta = -2,40$ si ottiene

$$Pr\{X_{\nu i} = 1\} = \frac{exp(0 + 2,40)}{1 + exp(0 + 2,40)} = 0,92$$

E per $\theta = 0$ e $\delta = +2,10$ si ottiene

$$Pr\{X_{\nu i} = 1\} = \frac{exp(0 - 2,10)}{1 + exp(0 - 2,10)} = 0,11$$

Ogni item del *SLM* ha una pendenza (*slope*) pari a 0,25, calcolata come derivata prima della *ICC* nel punto in cui $\theta = \delta$;

$$\begin{aligned}
\frac{\partial \pi_{\theta i}}{\partial \theta} &= \frac{\partial [e^{(\theta-\delta_i)}/(1 + e^{(\theta-\delta_i)})]}{\partial \theta} \\
&= (e^{(\theta-\delta_i)}) \frac{\partial}{\partial \theta} \frac{1}{(1 + e^{(\theta-\delta_i)})} + \frac{1}{(1 + e^{(\theta-\delta_i)})} \frac{\partial (e^{(\theta-\delta_i)})}{\partial \theta} \\
&= (e^{(\theta-\delta_i)})(-1) \frac{e^{(\theta-\delta_i)}}{(1 + e^{(\theta-\delta_i)})^2} + \frac{e^{(\theta-\delta_i)}}{(1 + e^{(\theta-\delta_i)})} \\
&= \frac{e^{(\theta-\delta_i)}}{(1 + e^{(\theta-\delta_i)})} - \frac{(e^{(\theta-\delta_i)})^2}{(1 + e^{(\theta-\delta_i)})^2} \\
&= \pi_{\theta i} - \pi_{\theta i}^2 = \pi_{\theta i}(1 - \pi_{\theta i}).
\end{aligned}$$

Se $\theta = \delta$, allora $\pi_{\theta i} = (1 - \pi_{\theta i}) = 0.5$ e quindi:

$$\frac{\partial \pi_{\theta i}}{\partial \theta} = \pi_{\theta i}(1 - \pi_{\theta i}) = (0.5)(0.5) = 0.25.$$

Si possono trarre valide indicazioni sulla bontà dell'item, cioè su come i dati si adattano al modello, osservando come si dispongono i punti attorno alla curva teorica. Quanto più i punti sono vicini alla curva tanto più si è portati a credere che i dati soddisfano gli assiomi del modello (monotonicità, *trait independence*, *response independence*, assenza di *DIF*).

La figura 4.6 rappresenta l'*ICC* della domanda 26. I valori medi delle quattro classi si collocano perfettamente sulla curva teorica stimata, pertanto si può supporre che l'item si adatti perfettamente al modello. Viceversa, i grafici 4.7 (item 10), 4.8 (item 13) e 4.9 (item 15) rappresentano le *ICC* di tre item che mal si adattano al modello (si è in presenza di *misfit* degli item); dal modo in cui le medie osservate si dispongo rispetto alle curve teoriche si può ipotizzare che questi tre item violino le ipotesi del modello, ognuno per ragioni diverse.

L'item 10 presenta, per le abilità più basse, le medie osservate al di sotto della curva teorica e, per le abilità più alte, le medie osservate al di sopra della curva teorica. Si è in presenza di un item che manifesta un'*over discrimination*, cioè di un item che discrimina tra le abilità degli studenti più di quanto richiesto dal modello. Come già ricordato in precedenza l'*over discrimination* il più delle volte è dovuta al fatto che l'item viola l'ipotesi di *response independence*. In questo caso la domanda, che presenta un coefficiente punto-biseriale molto alto (+0,525), come illustrato nella figura 4.30, è stata presa dalle prove del TIMSS e pone un quesito che richiede, oltre che la conoscenza delle frazioni, anche le capacità di riconoscere e interpretare l'uso delle frazioni nelle figure piane. Molto probabilmente questo item va a misurare capacità già contemplate da altri quesiti (item 2, 17, 19, 27 e 28). Uno studio sulla *local independence* potrebbe rivelare se ciò è vero.

L'item 13, viceversa, rappresenta un caso in cui le medie osservate si trovano al di sopra della curva teorica per i livelli bassi di abilità e al di sotto per i livelli alti di abilità. È un caso di *under discrimination* in cui l'item non coglie sufficientemente bene le differenze d'abilità tra gli studenti.

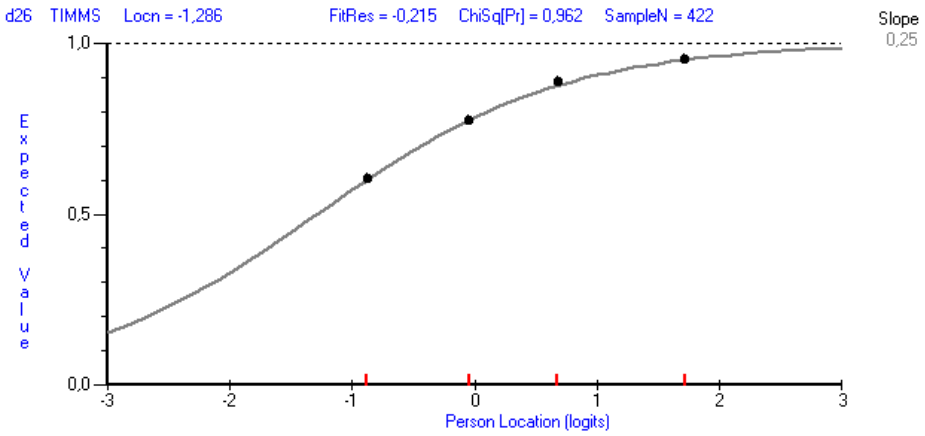


Figura 4.6: ICC dell'item d26

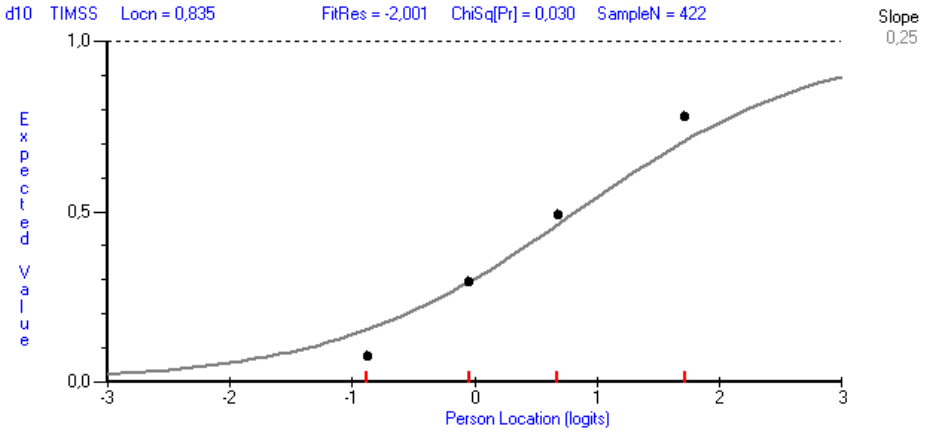


Figura 4.7: ICC dell'item d10

Il più delle volte l'*under discrimination* di un item è dovuta al fatto che esso coglie una caratteristica dell'individuo diversa o non altamente correlata con quella misurata dagli altri item (*multidimensionality*). A ben guardare, la domanda si riferisce alla conoscenza di una proprietà delle operazioni aritmetiche; nessun altro item del questionario considera questo campo della conoscenza e il saper riconoscere le proprietà delle operazioni può essere considerato ininfluente, o quasi, per la risoluzione degli altri quesiti.

L'interpretazione del grafico dell'item 15 è sicuramente più difficile. Qui si vede che la probabilità osservata è più alta della probabilità stimata per i livelli di abilità più bassi; diventa significativamente più bassa della curva teorica man mano che l'abilità cresce fino a posizionarsi sulla curva per i livelli più alti. L'item 15 è stato preso dai questionari del SNV per monitorare le classi V. È risultato mediamente molto difficile e, probabilmente, non tutte le classi hanno trattato l'argomento oggetto dell'item e solo gli scolari più brillanti hanno saputo rispondervi.

Dopo aver studiato le *ICC*, la validazione degli item avviene anche analizzando il comportamento dei residui, dal test del chi-quadrato e dal confronto tra le medie osservate e le medie attese. Nella tabella 4.10 per ogni item sono riportati, nell'ordine, il numero sequenziale, l'identificativo, il tipo, la posizione (difficoltà stimata), lo *standard error* della stima, il residuo coi suoi gradi di libertà, la statistica chi-quadrato coi suoi gradi di libertà e il *p-value*.

In base all'analisi del *fit* e all'osservazione delle *ICC* gli item che non si adattano al modello sono sei: il 5 (domanda 27 del questionario SNV 2004/2005), 10 (domanda del TIMSS 2007), 13 (domanda della prova SNV 2008/2009 per le classi V), 15 (domanda della prova SNV 2008/2009 per le classi V), 18 (domanda 21 del questionario SNV 2005/2006), e il 30 (domanda della prova SNV 2008/2009 per le classi V). Gli item 16 e 21, pur avendo un *p-value* piuttosto basso in partenza, migliorano il loro adattamento dopo la rimozione degli altri item, rientrando nei limiti di accettazione.

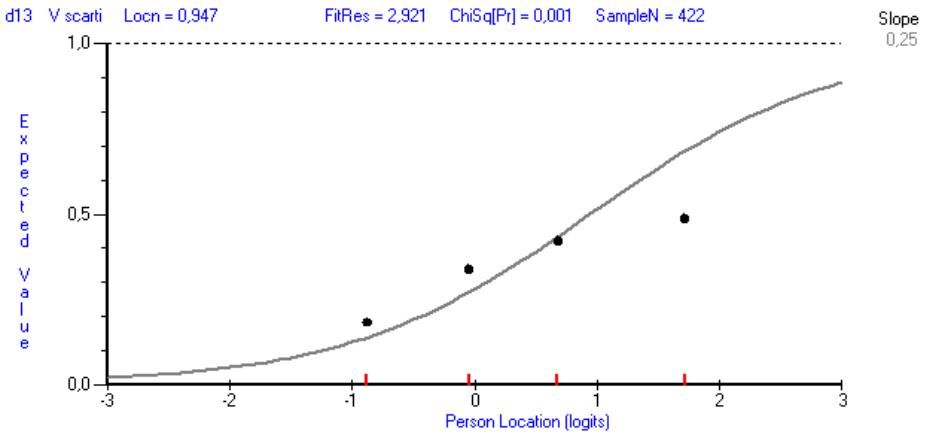


Figura 4.8: ICC dell'item d13

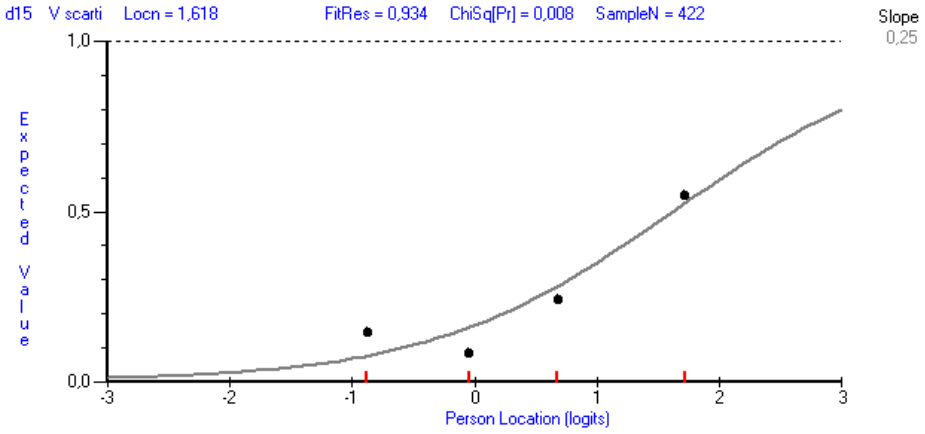


Figura 4.9: ICC dell'item d15

RUMM2020 Project: Test 2008/2009 Analysis: ANALISI1
Title: TUTTI
Display: INDIVIDUAL ITEM-FIT - Serial Order

Seq	Item	Type	Location	SE	Residual	DF	ChiSq	DF	Prob
1	d01	Poly	-2,402	0,180	-0,652	407,84	1,650	3	0,648115
2	d02	Poly	-1,396	0,131	0,644	407,84	5,600	3	0,132790
3	d03	Poly	-0,470	0,111	-0,792	407,84	5,353	3	0,147685
4	d04	Poly	-0,999	0,120	-0,744	407,84	1,369	3	0,712864
5	d05	Poly	0,528	0,107	2,547	407,84	6,592	3	0,086103
6	d06	Poly	0,390	0,107	1,217	407,84	3,989	3	0,262709
7	d07	Poly	-1,531	0,136	-0,453	407,84	2,558	3	0,464887
8	d08	Poly	-1,507	0,135	-0,693	407,84	1,701	3	0,636671
9	d09	Poly	0,234	0,107	-0,598	407,84	5,133	3	0,162342
10	d10	Poly	0,835	0,109	-2,001	407,84	8,930	3	0,030234
11	d11	Poly	0,513	0,107	0,185	407,84	7,977	3	0,046497
12	d12	Poly	-0,135	0,108	-0,597	407,84	4,143	3	0,246392
13	d13	Poly	0,947	0,111	2,921	407,84	17,183	3	0,000649
14	d14	Poly	-0,229	0,108	-0,316	407,84	1,582	3	0,663414
15	d15	Poly	1,618	0,123	0,934	407,84	11,853	3	0,007904
16	d16	Poly	-1,980	0,155	-1,439	407,84	8,693	3	0,033669
17	d17	Poly	0,606	0,108	2,120	407,84	6,145	3	0,104778
18	d18	Poly	0,488	0,107	-2,581	407,84	10,713	3	0,013383
19	d19	Poly	-0,239	0,109	-1,192	407,84	4,842	3	0,183709
20	d20	Poly	0,280	0,107	-0,910	407,84	1,689	3	0,639372
21	d21	Poly	0,384	0,107	-2,500	407,84	10,729	3	0,013287
22	d22	Poly	2,039	0,135	-0,213	407,84	0,366	3	0,947190
23	d23	Poly	0,695	0,108	0,769	407,84	1,130	3	0,769834
24	d24	Poly	-0,420	0,110	-0,008	407,84	3,541	3	0,315525
25	d25	Poly	0,452	0,107	-0,130	407,84	3,516	3	0,318735
26	d26	Poly	-1,286	0,128	-0,215	407,84	0,287	3	0,962380
27	d27	Poly	-0,553	0,112	-0,022	407,84	1,171	3	0,759895
28	d28	Poly	1,293	0,116	-0,543	407,84	3,916	3	0,270683
29	d29	Poly	-1,351	0,130	-0,781	407,84	3,087	3	0,378459
30	d30	Poly	1,924	0,131	2,921	407,84	40,442	3	0,000000
31	d31	Poly	2,098	0,137	1,018	407,84	5,730	3	0,125538
32	d32	Poly	-0,824	0,117	0,875	407,84	6,751	3	0,080260

Figura 4.10: Individual Item Fit test di link

Nelle figure 4.11 - 4.16 sono riportate le statistiche del test di adattamento item-tratto dei sei item che presentano un basso livello di adattamento. In ogni finestra compaiono, nell'ordine, il numero della classe intervallo, la sua numerosità, l'abilità stimata massima e l'abilità stimata media della classe, i residui dell'intervallo, cioè le differenze tra numero osservato e valore atteso di soggetti che hanno risposto in maniera corretta all'item, e la componente del chi-quadrato della classe (la somma delle componenti di tutte le classi danno come risultato il valore del chi-quadrato dell'item). A fianco, nella colonna *Score* sono riportate, per ogni categoria di risposta:

- le proporzioni osservate di risposte corrette e risposte errate (*OBS.P*);
- la proporzioni stimate di risposte corrette e risposte errate (*EST:P*);
- la probabilità che il risultato della risposta sia pari a 1, subordinatamente al fatto che esso possa essere o 0 o 1 (*OBS.T*). Nel caso di item dicotomici quest'ultima equivale alla proporzione osservata di risposte corrette; nel caso, più interessante, di item con più di 2 categorie di risposta (come nel *Partial Credit Model*) essa rappresenta la probabilità condizionata di rispondere nella categoria superiore di due categorie adiacenti.

Per ogni classe intervallo sono indicati i valori medi osservati (*OM*), i valori attesi (*EV*) e la loro differenza (*OM-EV*); anche in questo caso, trattandosi di item dicotomici, *OM* corrisponde a *OBS.P* e *EV* a *EST.P* delle risposte corrette. Infine, in basso, compare il valore del chi-quadrato riferito all'item, i suoi gradi di libertà e il suo *p-value*.

Si prenda in considerazione, a titolo di esempio, la figura 4.13 che riporta l'output dell'*Item-trait Chi-Square Test of Fit Statistics* per l'item d13. Coerentemente con quanto visto dall'esame della sua *ICC* per le due classi più piccole, i valori osservati sono superiori ai valori teorici (per la prima classe intervallo *OM-EV* = 0,04, per la seconda *OM-EV* = 0,06), per la terza classe i due valori sono quasi coincidenti (*OM-EV* = -0,01) mentre, per la quarta classe i valori osservati sono molto al di sotto dei valori teorici (*OM-EV* = -0,18). L'item evidenzia una marcata *under discrimination*, determinata per la maggior parte dall'ultima classe. Nelle quattro classi i residui (*Interval residual*)

valgono 1,096, 1,566, -0,248 e -3,670 rispettivamente mentre le componenti del chi-quadrato (*Chi-Square components*) valgono 1,201, 2,453, 0,062 e 13,467, rispettivamente. La somma delle *Chi-Square components* determina il valore della statistica chi-quadrato per l'item che vale 17,183 (con 3 gdl) e ha un *p-value* pari a 0,000649.
La figura 4.14 riporta l'output per l'item d15. In questo caso le differenze tra proporzioni osservate e proporzioni attese sono pari a 0,06, -0,07, -0,04 e 0,03; i residui nelle classi 2,407, -2,185, -0,999 e 0,534 e le *Chi-Square components* 5,794, 4,775, 0,999 e 0,285. Il valore del chi-quadrato a livello di item è pari a 11,853, con un *p-value* di 0,007904.

4.2.4 L'analisi dei distrattori

L'analisi dei distrattori cerca di cogliere informazioni aggiuntive sulla bontà delle domande del questionario che l'analisi condotta sugli item, una volta che sono stati dicotomizzati, non è più in grado di rilevare. Essa consiste principalmente nell'osservare le curve dei distrattori per individuare dei pattern di risposta anomali. Può capitare, per esempio, che i distrattori riproducano una struttura di risposta simile a quella di un item politomico del *Partial Credit Model*, nel qual caso sarebbe opportuno estendere le categorie di risposta e trasformare l'item da dicotomico a politomico con diversi livelli di correttezza. In RUMM2020 questo tipo di analisi risulta abbastanza semplice perché c'è un'opzione consente di rappresentare graficamente, per ogni classe intervallo, le proporzioni di soggetti che hanno scelto ognuna delle quattro alternative di risposta.

Prendiamo in considerazione i distrattori di quattro item che non si adattano al modello. L'item 5 (figura 4.17) ha due distrattori (le risposte B e D) le cui proporzioni vanno a 0 man mano che le abilità degli studenti crescono e un distrattore (la risposta C) che, invece, mantiene alta la probabilità di essere scelto come opzione anche dagli studenti più bravi.

L'item 10 (figura 4.18) ha due distrattori che non sono stati quasi mai scelti come risposta corretta dagli studenti e un distruttore (la risposta C) la cui probabilità cala con la stessa inclinazione con cui la probabilità della risposta

corretta cresce. I distrattori A e D si sono rivelati inutili in quanto troppo lontani dalla risposta corretta mentre il distrattore C si è rivelato efficace.

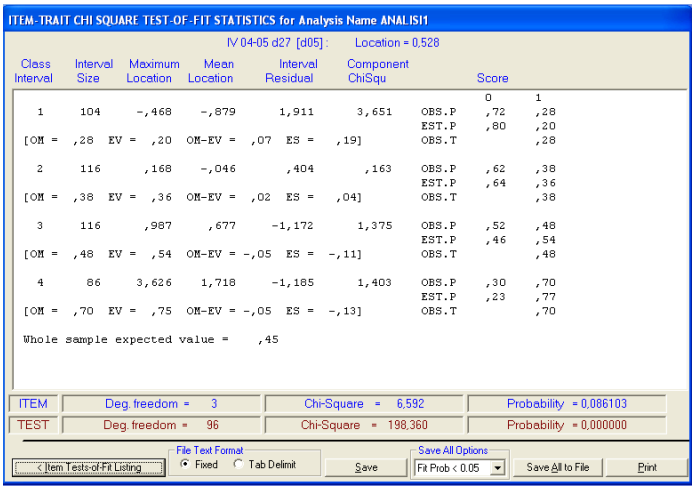


Figura 4.11: Test of Fit dell'item 5

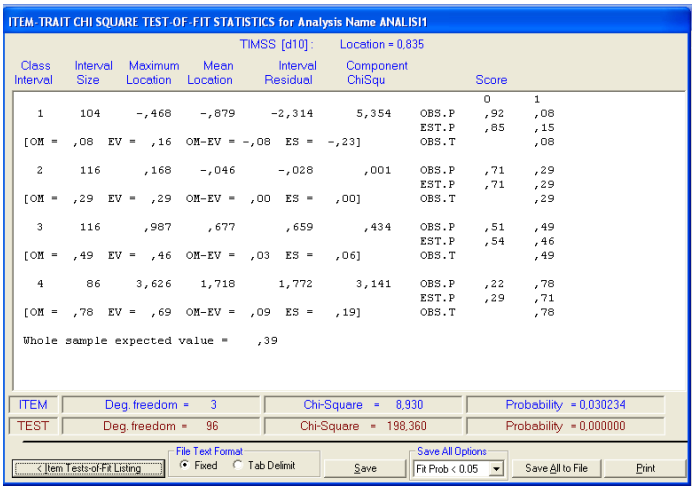


Figura 4.12: Test of Fit dell'item 10

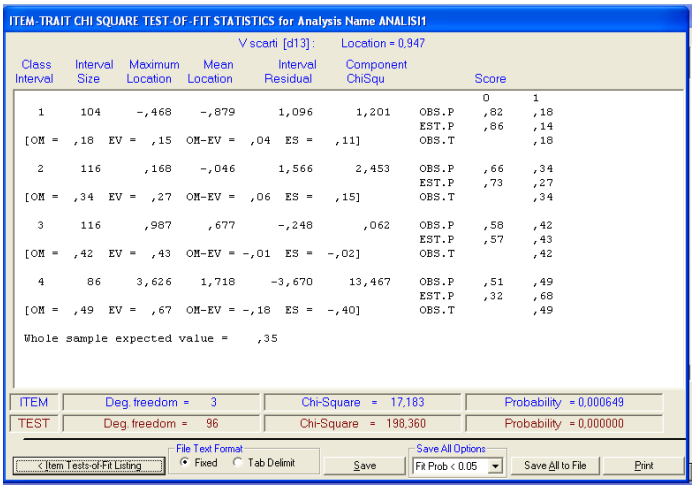


Figura 4.13: Test of Fit dell'item 13

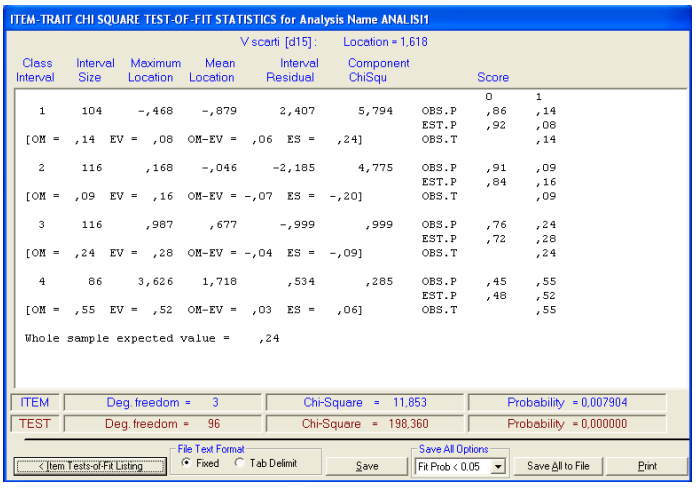


Figura 4.14: Test of Fit dell'item 15

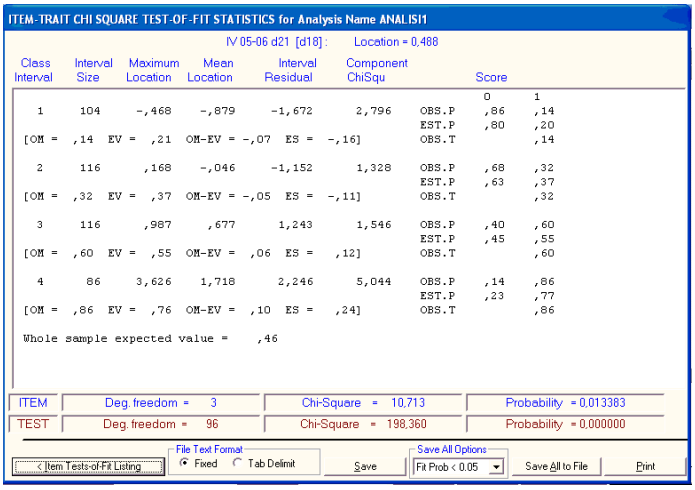


Figura 4.15: Test of Fit dell'item 18

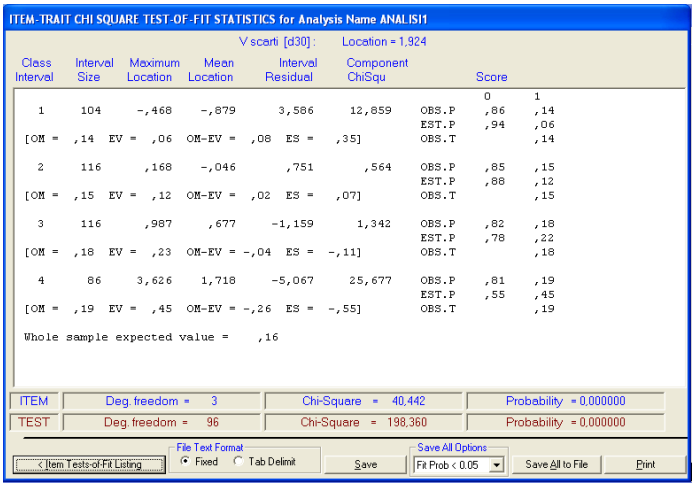


Figura 4.16: Test of Fit dell'item 30

L'item 15 (figura 4.19) ha un distrattore (la risposta B) che ha tratto in inganno quasi tutti gli studenti, eccezion fatta per quelli più bravi; infatti si vede che l'attrattiva esercitata da questa opzione è sempre maggiore di tutte le altre tre risposte (compresa quella corretta), salvo nell'ultimo tratto a destra della scala dove le probabilità osservate del distrattore scendono sotto le probabilità della risposta corretta.

L'item 30 (figura 4.20) è forse quello che ha il comportamento più anomalo. La curva di un distrattore (la risposta C) cala man mano che l'abilità cresce, la

curva di un altro distrattore (la risposta B) si mantiene praticamente costante lungo la scala e la curva di un distrattore (la risposta A) addirittura cresce con l'aumentare delle abilità .

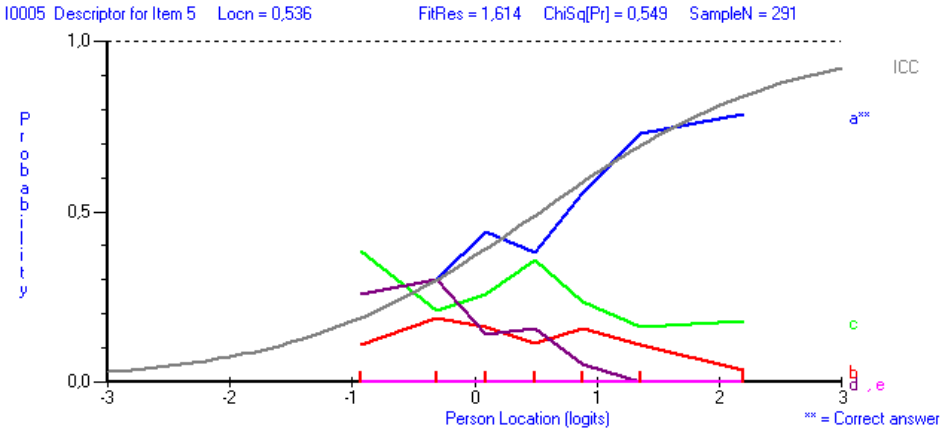


Figura 4.17: ICC dell'item 5 con le curve dei distrattori

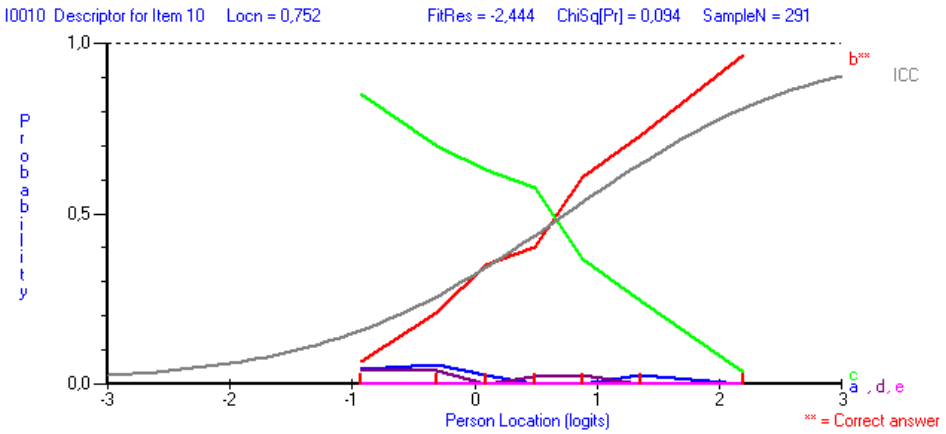


Figura 4.18: ICC dell'item 10 con le curve dei distrattori

4.2.5 L'analisi dei sottogruppi degli item

La Curva Caratteristica del Test (*Test Characteristic Curve*, in breve *TCC*) è la rappresentazione grafica della funzione che lega l'abilità degli individui con il punteggio ottenuto nel test. Va ricordato, infatti, che nel *RM* il punteggio

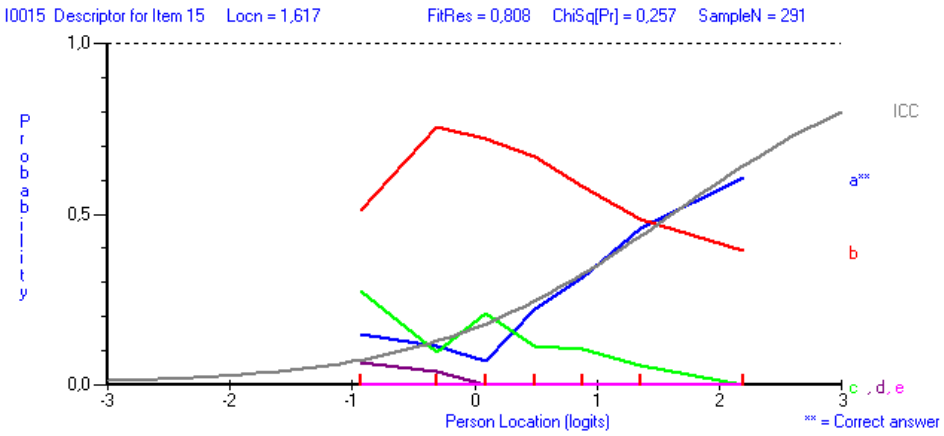


Figura 4.19: ICC dell'item 15 con le curve dei distrattori

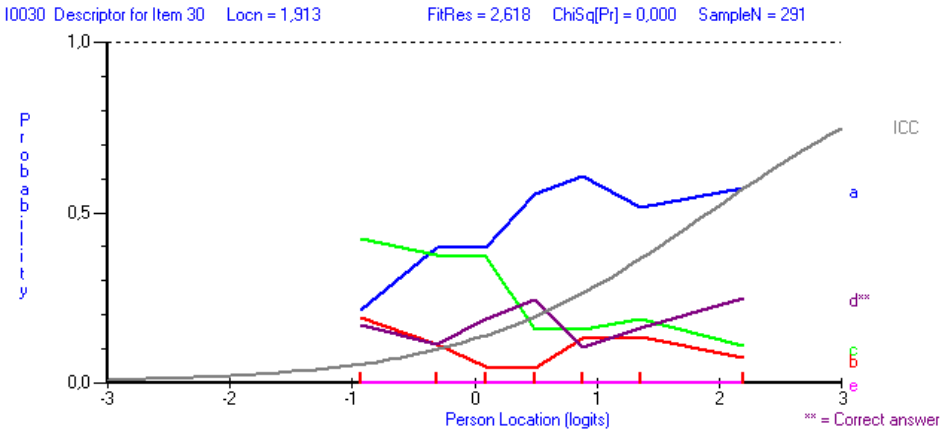


Figura 4.20: ICC dell'item 30 con le curve dei distrattori

ottenuto da ogni individuo (il suo *score*) costituisce la statistica sufficiente per stimare la sua abilità; a *score* uguali corrispondono abilità stimate uguali (nel caso in cui non siano presenti *missing values*, se, invece, ci sono dati mancanti, a *score* uguali possono corrispondere abilità diverse, come avviene nel caso di *link* di due o più test in cui a *score* uguali di test diversi corrispondono abilità differenti; questa eventualità sarà presa in esame nel prosieguo di tempo quando verrà trattato l'aggancio dei test SNV), indipendentemente da quali sono stati gli item a cui è stata data una risposta corretta. La *TCC* indica semplicemente quale punteggio ci si aspetta da un certo livello di abilità o quale livello

di abilità viene determinato da un determinato punteggio. Non dice nulla sulla bontà di adattamento dei dati o sulla qualità delle stime che vengono rilevate a livello di singoli item e di singoli individui mediante l'analisi dei residui però risulta uno strumento molto utile per confrontare l'equivalenza degli *score* in sottogruppi di item appartenenti allo stesso test o degli *score* realizzati in test differenti. Sempre rimanendo sul test di *link*, dall'*Item-Person map* della figura 4.24, si vede che il gruppo delle domande prese dal questionario di V è risultato mediamente più difficile del gruppo di domande del TIMSS che, a sua volta, lo è stato rispetto al gruppo di domande desunte dai questionari SNV somministrati negli anni passati alle IV.

La figura 4.21 rappresenta le Curve Caratteristiche (e le rispettive funzioni di Informazione) di tre gruppi di item distinti del test di *link*: il primo gruppo è formato dalle sei domande più difficili (item 13, 15, 21, 23, 30 e 31) tra quelle prese dal test SNV 2008/2009 per le classi V, il secondo gruppo è formato dalle sei domande più difficili del TIMSS (item 9, 10, 11, 20, 22, 28) e il terzo dalle sei domande più difficili tra quelle somministrate alle classi quarte nei vecchi questionari del SNV (item 3, 5, 14, 17, 18 e 25). Nel riquadro in alto a destra compare la tabella di equivalenza tra il punteggio ottenuto e le abilità stimate in ciascuno dei tre gruppi; ad esempio per ottenere uno score uguale a 3 nel sottogruppo di item del SNV 2008/2009 è necessaria un'abilità pari a 1,28, più alta di quella necessaria (0,84) per ottenere lo stesso *score* nel primo sottogruppo dei quesiti del TIMSS e di quella necessaria (0,24) per ottenere lo stesso *score* nel sottogruppo degli item delle IV. La figura 4.22 rappresenta la conversione grafica tra abilità e punteggi (*Location to Score*) nei tre test (nella fattispecie è indicata la corrispondenza degli *score tests* 2,5, 3,2 e 4,2 a un'abilità pari a 1) mentre la figura 4.23 la conversione grafica tra punteggi e abilità (*Score to Location*, nella fattispecie è indicata la corrispondenza delle abilità stimate 1,942, 1,508 e 0,854 a un *test score* pari a 4).

4.2.6 Difficoltà presunta e difficoltà stimata degli item

La stima dei parametri degli item ha permesso di accertare se la difficoltà attribuita "a priori" a un item coincide con la difficoltà stimata "a posteriori"

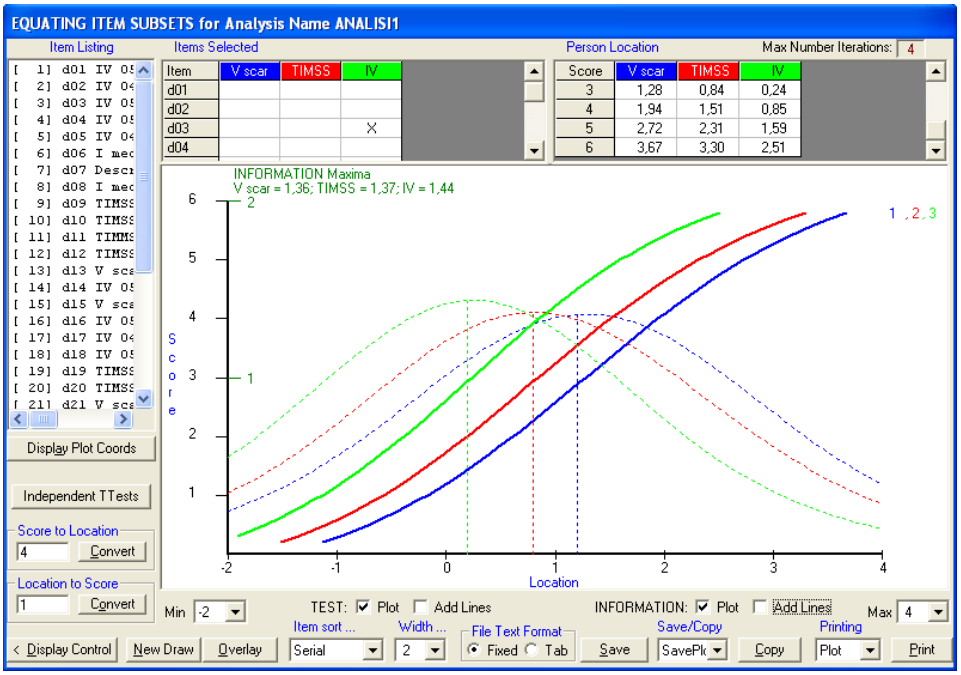


Figura 4.21: Test Characteristic Curves dei tre gruppi di item (SNV IV, SNV V, TIMSS)

dal modello. Capita abbastanza spesso, infatti, soprattutto quando i questionari vengono costruiti da persone che non sono gli insegnanti degli studenti a cui viene somministrato il test, che domande ritenute molto impegnative risultino relativamente semplici e che, viceversa, quesiti di cui si presuppone un'alta percentuale di risposte corrette abbiano poi uno scarso grado di successo. Tale fatto costituisce una delle ragioni principali per cui è sbagliato assegnare ad item pesi diversi nell'assegnazione dei punteggi (voti) agli studenti. Nei modelli di Rasch tutti gli item hanno lo stesso peso; poiché lo *score* ottenuto da un individuo costituisce una statistica sufficiente per la stima del suo livello di abilità, due persone diverse che hanno ottenuto lo stesso *score* risultano parimenti abili, nonostante possano aver risposto, come ipotesi limite, uno a tutti gli item più facili e uno a tutti gli item più difficili. La figura 4.24 rappresenta gli item collocati sulla scala di misura in base alla loro difficoltà e indica per ciascuno di essi, tra parentesi, il livello presunto di difficoltà e la tipologia di tema a cui appartiene. È facile osservare che alcuni item ritenuti facili si collo-

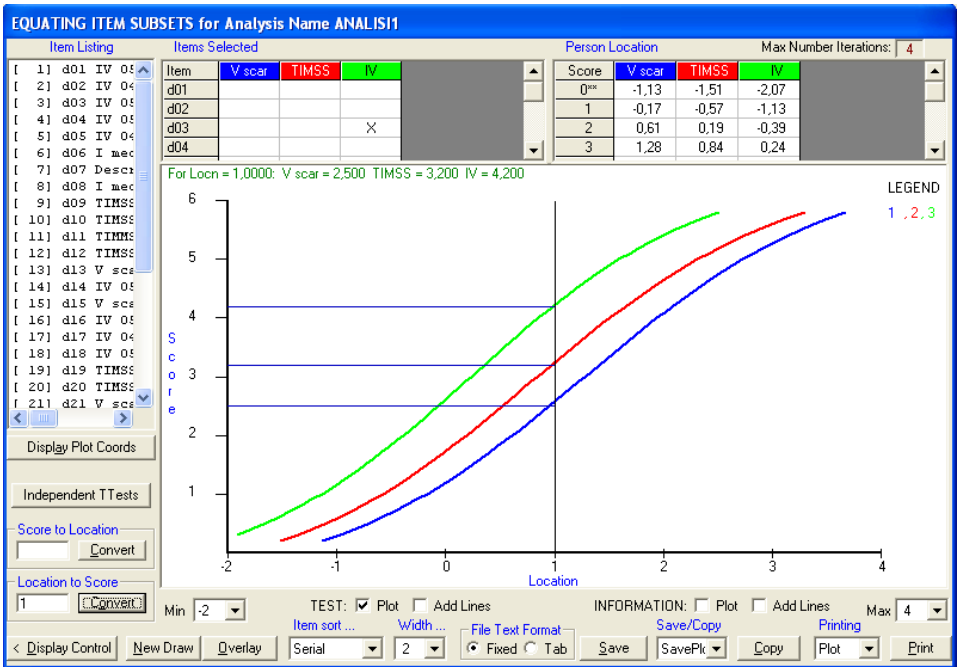


Figura 4.22: Test Characteristic Curves dei tre gruppi di item: location to score

cano nella parte medio alta della scala e che item ritenuti più difficili occupano livelli bassi della stessa.

4.2.7 L'analisi degli studenti

Le statistiche di sintesi contenute nella finestra *Summary Statistics* forniscono le prime indicazioni di massima riguardanti le stime delle abilità degli alunni cui è stato somministrato il test e i relativi residui. Sulla base dei dati analizzati si può affermare che il test è stato leggermente facile per gli scolari del campione ma che, tuttavia, è riuscito a distinguere soddisfacentemente bene tra le loro abilità. Come già detto l'analisi generale dei residui indica una distribuzione che si discosta dall'ipotetica normale standardizzata in quanto manifesta una moderata curtosi negativa e un'asimmetria positiva.

L'analisi individuale dei residui ha l'obbiettivo di determinare, per ogni soggetto, lo scostamento del suo *pattern* di risposte da quello atteso dal modello. Bisogna tener presente che, una volta che le abilità dei soggetti e le

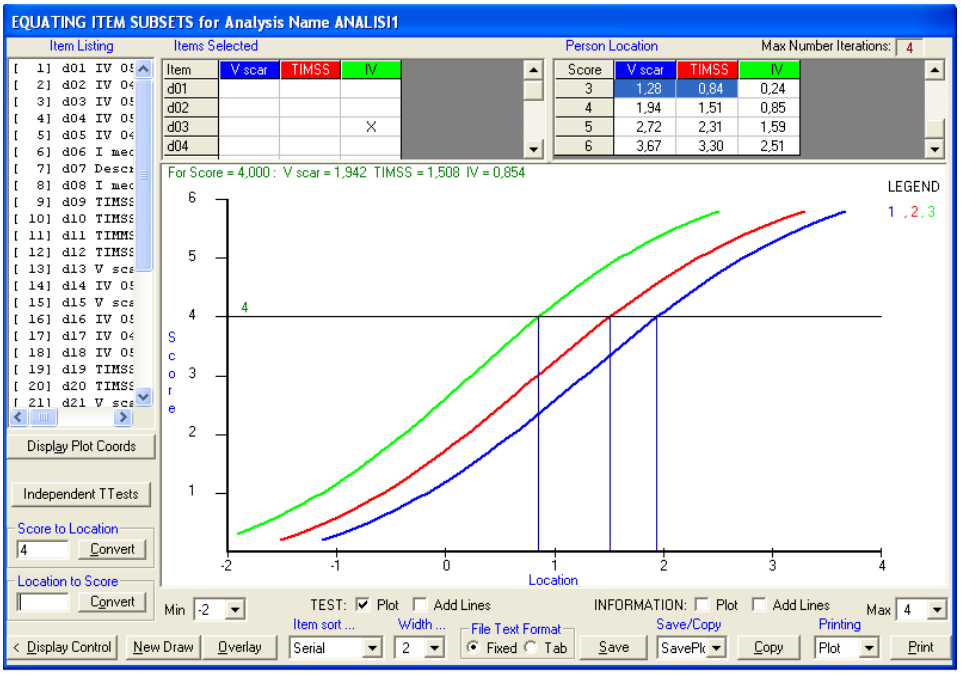


Figura 4.23: Test Characteristic Curves dei tre gruppi di item: score to location

difficoltà degli item sono state stimate e ordinate, la matrice a scala di Guttman è, tra tutte le matrici n x K osservabili, quella che ha la probabilità più alta di verificarsi. Il confronto fra *pattern* di risposta e *pattern* attesi avviene, quindi, tenendo come riferimento il *pattern* di Guttman, nel senso che il *pattern* di risposta di Guttman è quello più vicino al *pattern* teorico del modello, mentre il *pattern* di risposta anti-Guttman è quello che si allontana di più dal *pattern* di risposta teorico. Nel RM, però, un *pattern* di risposta troppo vicino al *pattern* di Guttman è comunque indice di un cattivo adattamento al modello, in quanto manifestazione di una struttura deterministica nelle risposte, imputabile a una violazione della *response independence*. Tradotto in termini di analisi dei residui, ciò significa che valori sia troppo alti che troppo bassi di *Fit residual* rappresentano una discordanza i dati e il modello. La figura 4.25 riporta una porzione della finestra di output dell'*Individual Person Fit*. Essa contiene, oltre alle informazioni generali di adattamento, riportate in basso, il numero seriale di ogni soggetto (*recID*), il suo punteggio ottenuto (*Totsc*, che nel caso dicotomico in questione equivale al numero di risposte corrette

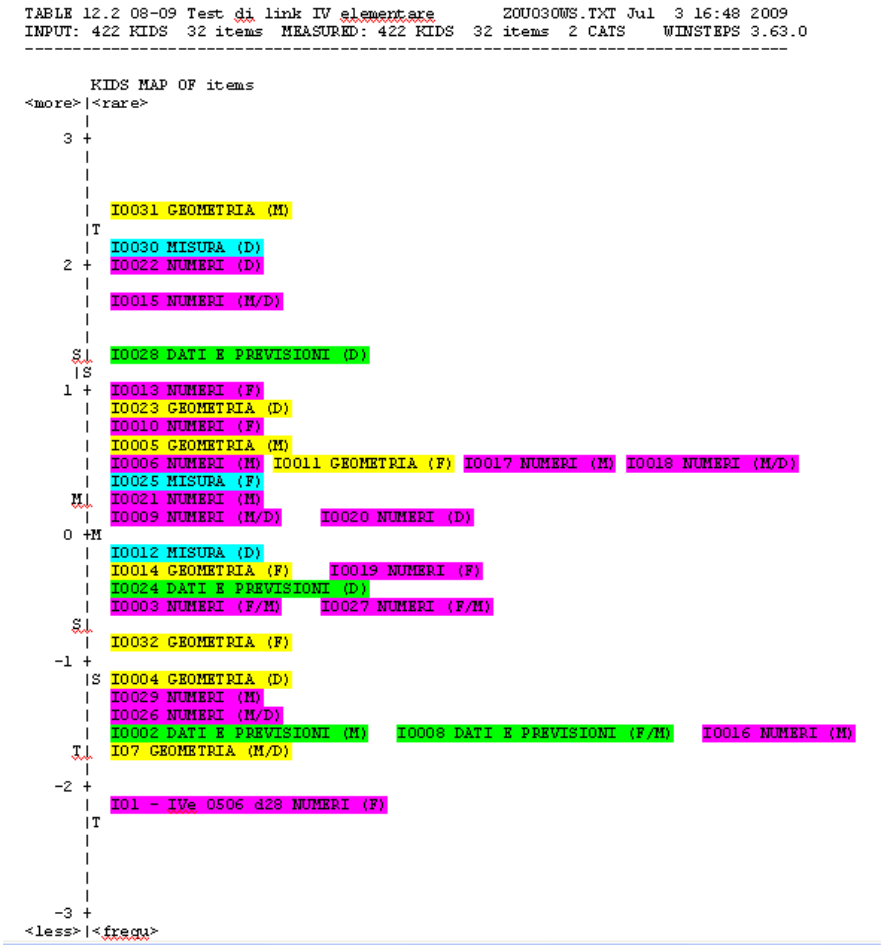


Figura 4.24: Temi e difficolta

date), il numero di domande a cui ha dato una risposta (*Maxsc*), il numero degli item che gli sono stati somministrati (*item*), un indicatore per segnalare i punteggi estremi (*Extreme*), cioè gli alunni che hanno risposto correttamente o incorrettamente a tutte le domande (nella fattispecie non ci sono casi estremi, il punteggio minimo realizzato è 4 e quello massimo 31), la stima dell'abilità del soggetto lungo il continuum (*Location*), l'errore di misura (*SE*), il residuo (*FitResid*), i gradi di libertà (*DegFree*) e l'identificativo (*ID*).

La figura 4.26 contiene l'output dell'*Individual Person Fit* degli scolari con il più basso e il più alto valore di *Fit residual*. Il soggetto 120, ad esempio, ha un Residuo molto basso (-2,298) che indica che il suo *pattern* di risposta

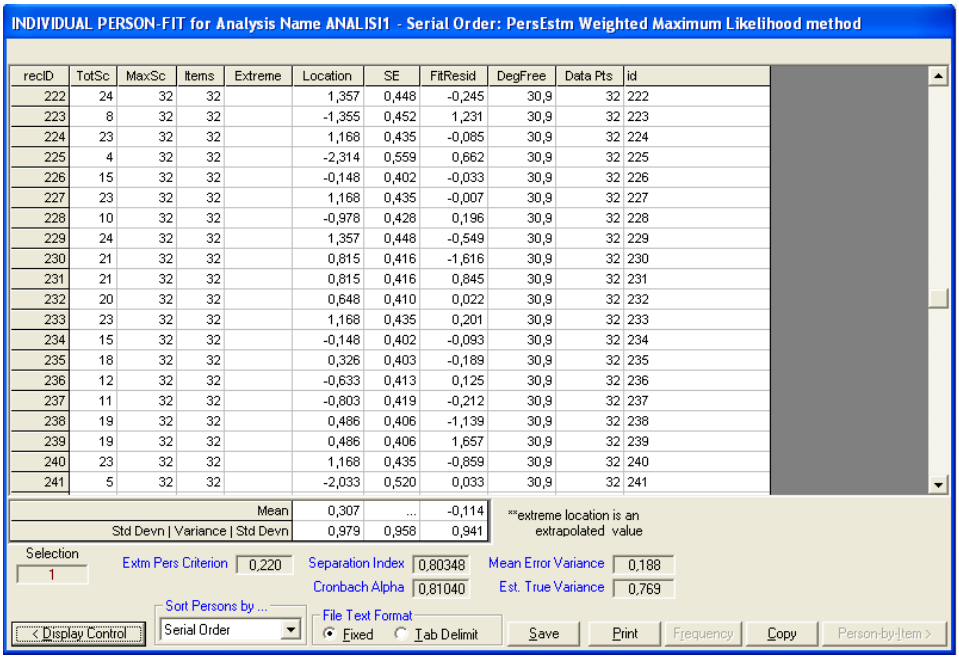


Figura 4.25: Individual Person Fit

riproduce troppo bene i valori attesi del modello. Osserviamo la sua stringa di risposte riportata nella tabella 4.2. Egli ha risposto correttamente ai dodici item più facili, ha sbagliato l'item 19 e l'item 14, ha risposto correttamente all'item 12, ha sbagliato il 9, ha risposto correttamente al 20 e ha sbagliato tutti gli item successivi, ad eccezione del 23. Il suo *pattern* di risposta è molto simile al pattern di Guttman in cui, a una sequenza di tutti 1 per le domande più facili, segue una sequenza di tutti 0 per le domande più difficili. Viceversa, il soggetto 145 è quello con il *Fit residual* più alto (+2,478). La sua stringa di risposta è l'ultima della tabella 4.2; alla prima risposta corretta seguono tre risposte sbagliate, poi un'altra corretta e altre due sbagliate, e così avanti fino all'ultimo item (il più difficile) a cui ha risposto correttamente. Il *pattern* è molto dissimile da quello di Guttman: le risposte corrette si alternano a quelle errate lungo tutta la scala delle difficoltà degli item come se le risposte fossero state date a caso. Mentre un pattern di risposta molto simile al pattern di Guttman è facile da giustificare anche in via teorica, un pattern con un elevato residuo per essere spiegato richiede un'indagine "ad hoc" più approfondita in quanto può celare problemi a livello di singolo individuo (*cheating*, *guessing*),

ID	Total	Max	Miss	Extreme	Locn	SE	Residual	DegFree	DataPts	id
120	15	32	32		-0,148	0,40	-2,298	30,90	32	120
101	15	32	32		-0,148	0,40	-2,259	30,90	32	101
148	14	32	32		-0,307	0,40	-2,219	30,90	32	148
398	18	32	32		0,326	0,40	-1,977	30,90	32	398
355	22	32	32		0,987	0,42	-1,884	30,90	32	355
294	23	32	32		1,168	0,44	-1,852	30,90	32	294
166	19	32	32		0,486	0,41	-1,823	30,90	32	166
416	13	32	32		-0,468	0,41	-1,783	30,90	32	416
255	14	32	32		-0,307	0,40	-1,698	30,90	32	255
284	24	32	32		1,357	0,45	-1,696	30,90	32	284
...	...	...	...		...	...	...	...	...	...
409	12	32	32		-0,633	0,41	1,714	30,90	32	409
46	20	32	32		0,648	0,41	1,719	30,90	32	046
19	20	32	32		0,648	0,41	1,724	30,90	32	019
380	17	32	32		0,168	0,40	1,781	30,90	32	380
65	17	32	32		0,168	0,40	1,782	30,90	32	065
361	9	32	32		-1,162	0,44	1,806	30,90	32	361
319	12	32	32		-0,633	0,41	1,837	30,90	32	319
124	12	32	32		-0,633	0,41	1,901	30,90	32	124
420	10	32	32		-0,978	0,43	1,924	30,90	32	420
35	21	32	32		0,815	0,42	2,004	30,90	32	035
270	17	32	32		0,168	0,40	2,064	30,90	32	270
187	18	32	32		0,326	0,40	2,093	30,90	32	187
32	10	32	32		-0,978	0,43	2,249	30,90	32	032
145	11	32	32		-0,803	0,42	2,478	30,90	32	145
				Mean:	0,307		-0,114			
				SD :	0,979		0,941			

Figura 4.26: I Person Fit più piccoli e più grandi

a livello di classe (argomenti non trattati dall'insegnante) o a livello di item (*misfitting* per un determinato gruppo di individui).

Item	Diffic.	120	101	148	187	32	145
d01	-2,4	1	1	1	1	1	1
d16	-1,98	1	1	1	0	0	0
d07	-1,53	1	1	1	1	1	0
d08	-1,51	1	1	1	1	0	0
d02	-1,4	1	1	1	1	1	0
d29	-1,35	1	1	1	1	0	1
d26	-1,29	1	1	1	0	0	0
d04	-1	1	1	1	1	1	0
d32	-0,82	1	0	1	1	1	1
d27	-0,55	1	1	1	0	0	1
d03	-0,47	1	1	0	1	0	0
d24	-0,42	1	1	1	0	1	1
d19	-0,24	0	1	1	1	0	0
d14	-0,23	0	0	0	0	0	1
d12	-0,13	1	1	0	0	0	0
d09	0,23	0	1	0	0	0	0
d20	0,28	1	1	0	1	1	0
d21	0,38	0	0	1	0	0	0
d06	0,39	0	0	1	1	0	1
d25	0,45	0	0	0	1	1	0
d18	0,49	0	0	0	0	0	0
d11	0,51	0	0	0	0	0	1
d05	0,53	0	0	0	1	0	0
d17	0,61	0	0	0	0	0	1
d23	0,69	1	0	0	1	0	0
d10	0,83	0	0	0	1	0	0
d13	0,95	0	0	0	0	0	0
d28	1,29	0	0	0	0	0	0
d15	1,62	0	0	0	1	0	0
d30	1,92	0	0	0	1	0	1
d22	2,04	0	0	0	0	1	0
d31	2,1	0	0	0	1	1	1
Id		120	101	148	187	32	145
Location		-0,148	-0,148	-0,307	0,326	-0,978	-0,803
Residual		-2,298	-2,259	-2,219	2,093	2,249	2,478

Tabella 4.2: Pattern di risposte, Location e Fit Residual

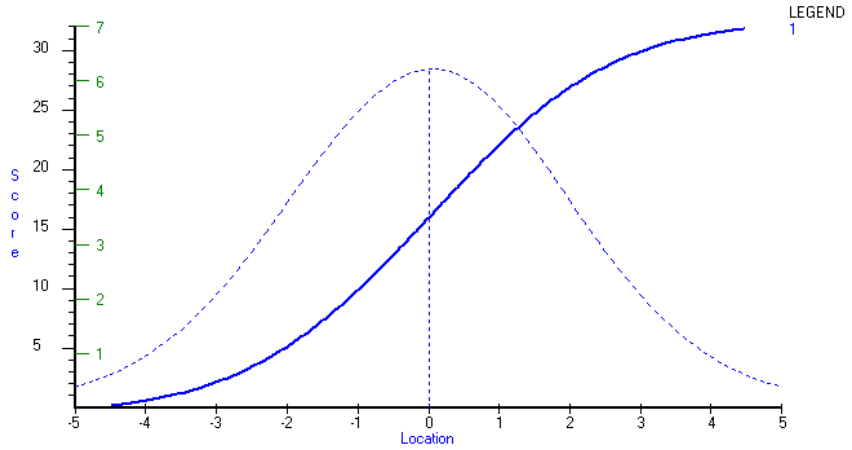


Figura 4.27: La Curva Caratteristica del Test

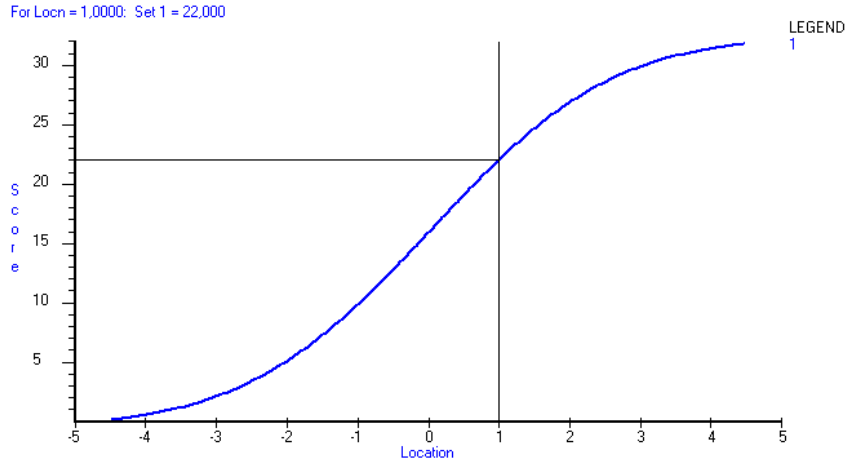


Figura 4.28: La Curva Caratteristica del Test: Location to score

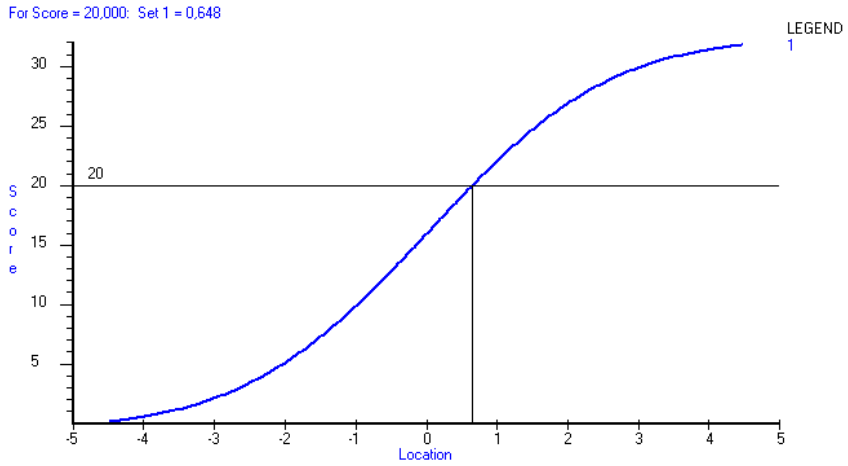


Figura 4.29: La Curva Caratteristica del Test: Score to location

4.2.8 Le statistiche dell'analisi classica dei test

RUMM2020 e Winsteps consentono anche di ottenere degli output con alcune statistiche relative alle analisi tradizionali sugli individui e sugli item. Per quanto riguarda gli individui, i software determinano le statistiche sufficienti e altre informazioni quali il numero di domande che sono state somministrate, il punteggio massimo ottenuto, le frequenze, assolute e cumulate, di coloro che hanno risposto a un determinato item. Per quanto riguarda gli item (figura 4.30) essi determinano il coefficiente di correlazione punto bi-seriale e l'Indice di Riproducibilità di Green. Il coefficiente di correlazione punto bi-seriale è utile per diagnosticare *miscoding* o malfunzionamento di un item; se il valore del coefficiente è negativo o pari a zero significa o che le risposte sono state codificate male o che l'item contrasta con la costruzione della variabile unidimensionale da misurare, se il valore è troppo elevato (contrariamente a quanto avviene nella Teoria Classica dei test in cui più è alta la correlazione migliore è l'item) è possibile ipotizzare una violazione della *response independece*.

Il coefficiente di correlazione punto-biseriale, indicato spesso con r_{pb}, è un caso particolare del coefficiente di correlazione di Pearson in cui una variabile

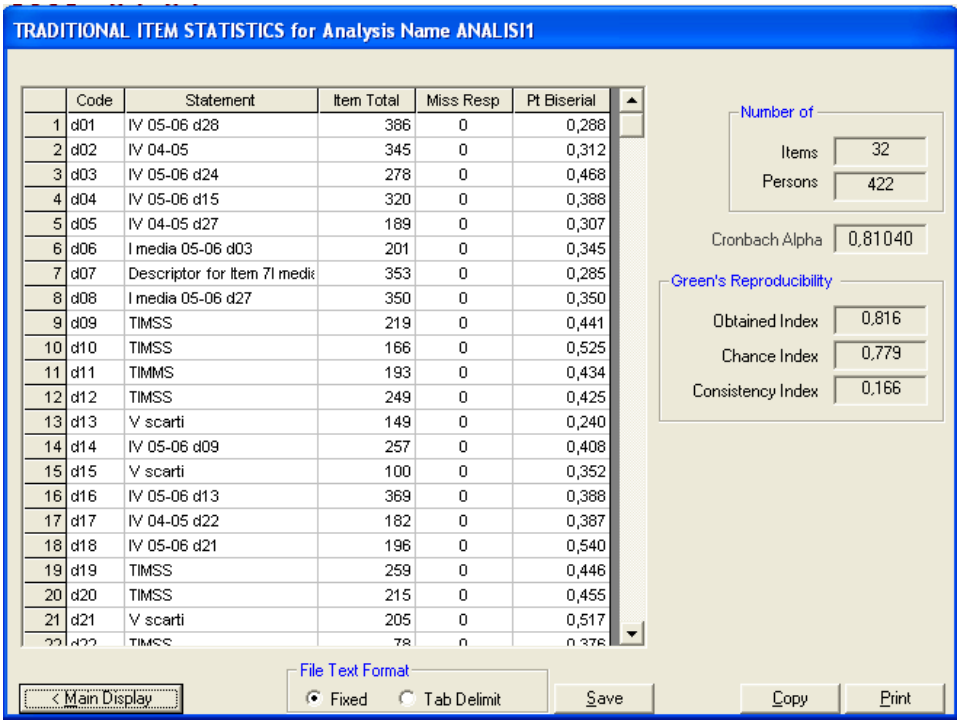


Figura 4.30: Traditional Item Statistics

è quantitativa e l'altra variabile è dicotomica. La formula finale è

$$r_{pb} = \frac{(\hat{Y}_1 - \hat{Y}_0)(pq)^{\frac{1}{2}}}{\sigma_Y} \tag{4.7}$$

dove $\hat{Y}_0$ e $\hat{Y}_1$ sono i punteggi medi di Y nei casi in cui $X = 0$ e $X = 1$, rispettivamente, p e q sono le due proporzioni dei dati in cui $X = 1$ e $X = 0$ e σ_Y è la deviazione standard di Y nella popolazione o nel campione osservato. Nella fattispecie r_{pb} indica l'associazione esistente tra le stime delle abilità e le risposte dicotomiche di ciascun item.

La matrice di Guttman (figura 4.31) riproduce il *data set* delle risposte evidenziando quelle corrette e quelle sbagliate; se tale *data set* è ordinato in base agli *score* delle persone e degli item si dice che esso riproduce un *pattern* di Guttman (secondo lo psicometrico Louis Guttman il test ideale era quello in cui ogni individuo rispondeva esattamente a tutti gli item fino a una certa difficoltà e poi sbagliava tutti quelli successivi; il numero di risposte corrette

rappresentava il suo livello di abilità). Come già sottolineato, il *pattern* di Guttman rispecchia una struttura deterministica delle risposte che si allontana dalla filosofia che sta alla base del RM. Una struttura deterministica, infatti, per Rasch indica una ridondanza di informazione apportata dai singoli item che si manifesta una violazione dell'assunzione di indipendenza locale e come tale va rifiutata. Inoltre una matrice di Guttman non permette la stima dei parametri in quanto ogni caso (individuo o item) risulta un caso estremo.

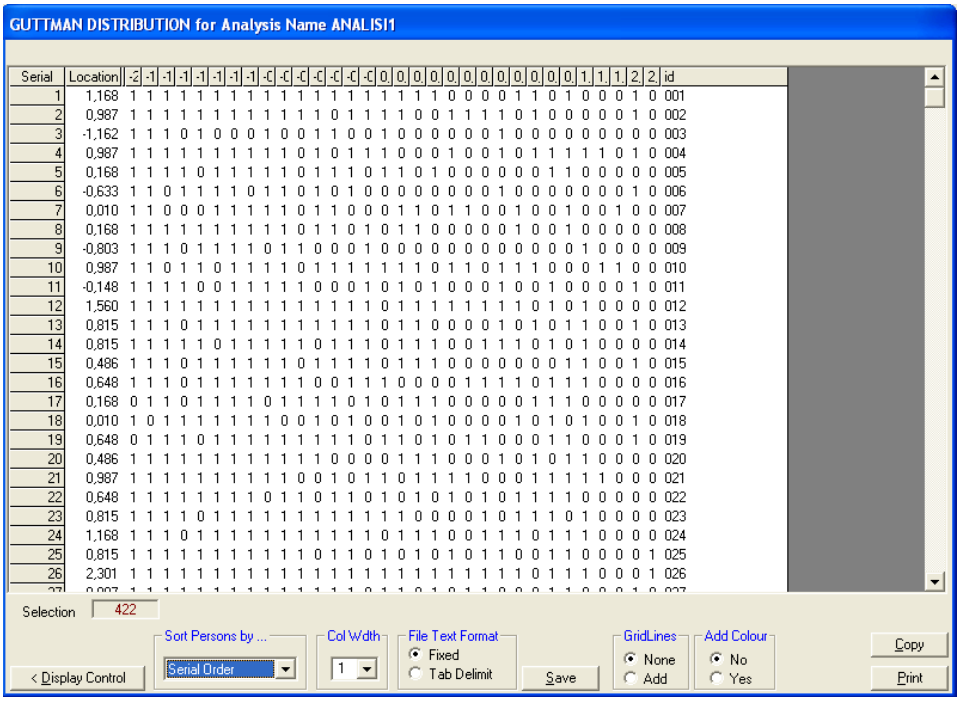


Figura 4.31: Guttman Distribution

Il *data set* visualizzato in forma di matrice di Guttman risulta però utile per analizzare i singoli pattern di risposta degli individui e identificare quelli che si avvicinano troppo alla struttura deterministica dei dati e quelli che, invece, rispecchiano una struttura antitetica a quella di Guttman, manifestando un alto grado di *misfit*. Sia le persone che gli item devono essere ordinati in ordine crescente in base alla loro rispettiva *location* prima di poter valutare la conformità dei dati alla matrice di Guttman.

L'indice di riproducibilità di Green consente di fare una valutazione alternativa del grado di conformità dei dati al *pattern* di Guttman.

4.2.9 Le Category Probability Curves e le Thresholds Probability Curves

Nella sezione di RUMM2020 dedicata all'analisi degli item è possibile specificare, oltre alle *Item Chararcteristic Curves* e alle *Distractor Curves* anche le *Category Probability Curves* (*CPC*) e le *Thresholds Probability Curves* (*TPC*). Queste ultime due rappresentazioni grafiche svolgono un importante ruolo nella valutazione visiva del comportamento delle stime delle soglie delle categorie di risposta nel caso venga adottato un *Rating Scale Model* o un *Partial Credit Model* mentre non apportano nessuna informazione aggiuntiva nel caso di *SLM*.

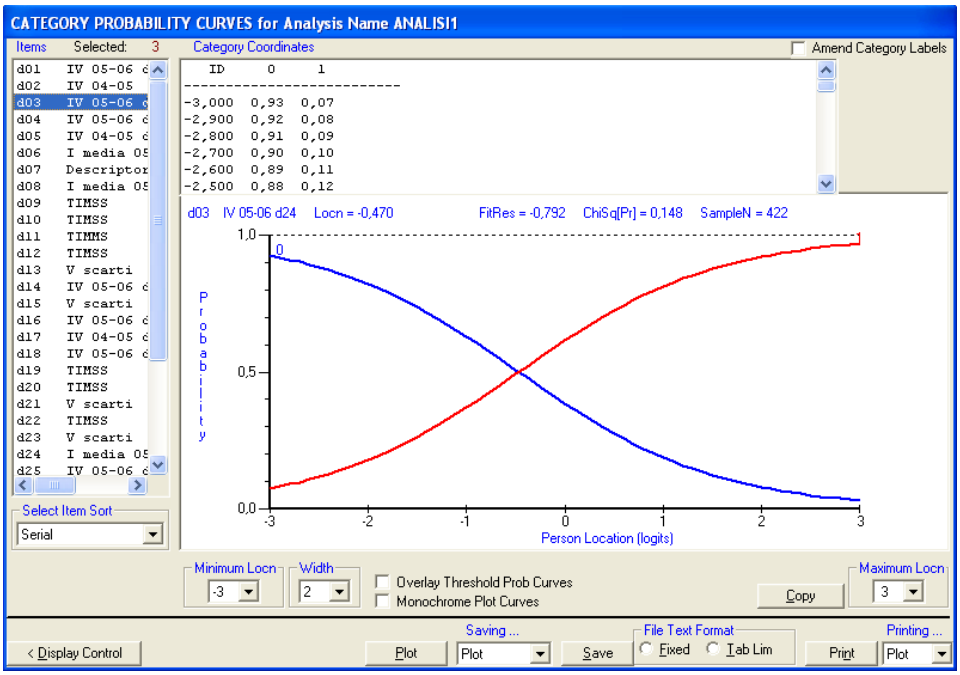


Figura 4.32: Category Probability Curves nel caso dicotomico

Le *Category Probability Curves* rappresentano le probabilità assolute di ciascuna categoria di risposta di ciascun item. I valori sul *continuum* dei punti in cui le curve si intersecano rappresentano le soglie delle categorie; nel caso dicotomico (figura 4.32) l'unica soglia, che separa la categoria delle risposte sbagliate dalla categoria delle risposte corrette, coincide con la difficoltà del-

l'item; nel caso politomico le soglie (figura 4.33) sono i valori stimati delle τ_k dell'ELM trattato nel paragrafo 2.2.3.

Le *Thresholds Probability Curves* rappresentano le curve di probabilità di ogni soglia. Nel caso del *SLM* esse coincidono con le *ICC*, mentre nel caso di item politomici esse rappresentano le curve delle probabilità condizionate di raggiungere il livello k anziché il livello k -1 espresse dalla 2.32. I valori sul continuum dei punti in cui le curve valgono 0,5 rappresentano le stime dei parametri δ_k del *PCM*, cioè i livelli di difficoltà delle varie categorie di risposta. Sia le *CPC* che le *TPC* servono a verificare il buon adattamento delle categorie degli item al modello. Infatti, se le categorie degli item funzionano bene l'ordine delle stime delle soglie delle categorie di risposta deve rispecchiare quello teorico del modello e le soglie (e di conseguenza i livelli di difficoltà delle varie categorie) devono essere ben distanziate tra loro in maniera che ogni categoria rappresenti in modo chiaro e definito un livello diverso di abilità. Le figure 4.33 e 4.34 rappresentano un item in cui le categorie di risposta sono ben definite e calibrate; le soglie rispettano l'ordinamento crescente, sono sufficientemente separate tra loro così come lo sono le curve di probabilità dei *thresholds*.

Per completezza viene riportato un caso (preso dagli esempi di RUMM2020) in cui i dati evidenziano un cattivo funzionamento delle categorie (figure 4.35 e 4.36). Si può osservare, innanzi tutto, che l'ordinamento delle soglie non è soddisfatto, risultando τ_3 minore di τ_2; in aggiunta, tutti i livelli delle soglie sono vicini uno all'altro sulla scala di misurazione. Da questi risultati si può dedurre che le categorie dell'item non discriminano correttamente tra le diverse abilità delle persone e che, anzi, la probabilità di selezionare la categoria 3 è quasi sempre superiore alla probabilità di selezionare la categoria 2, il che si manifesta chiaramente nel grafico delle TPC dove la curva della soglia 3 risulta più a sinistra delle curve delle soglie 1 e 2. Un risultato anomalo di questo tipo richiede una revisione accurata dell'item e delle categorie di risposta assegnate se non addirittura la sua rimozione dal modello.

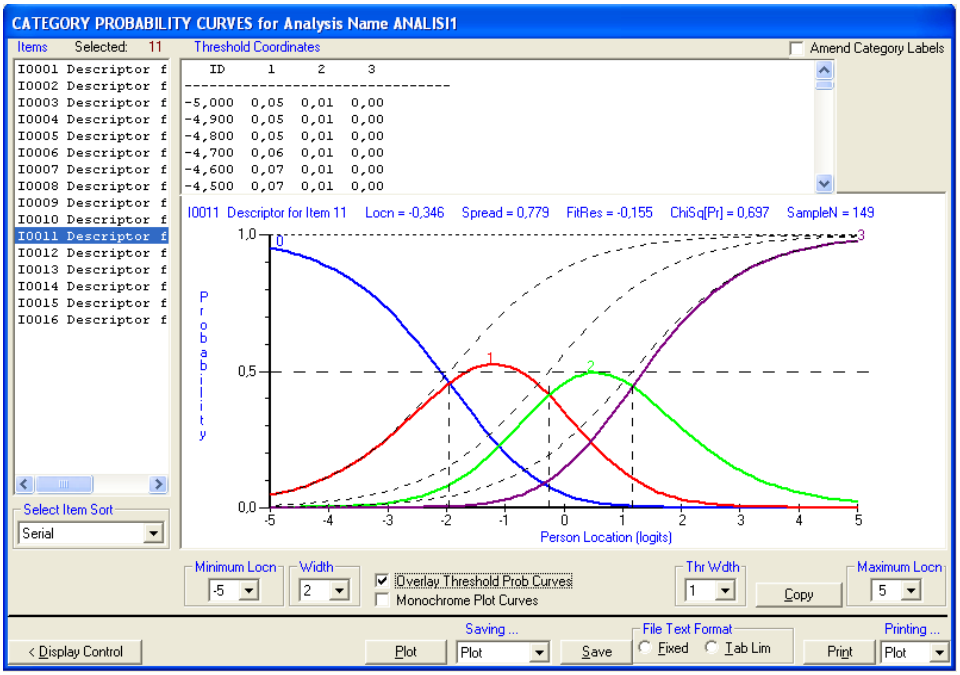


Figura 4.33: Category Probability Curves nel caso politomico

4.2.10 La pulizia del test

Alla luce dei risultati delle analisi esposte nei paragrafi precedenti relative all'adattamento dei dati al modello si è proceduto alla pulizia dei dati del test di link; si sono cioè eliminati gli item che hanno evidenziato il più alto *misfit* in termini di *Fit residual* e di soglia di rifiuto, nella fattispecie gli item 5, 10, 13, 15, 18 e 30 e si sono stimati nuovamente tutti i parametri del modello. Le nuove statistiche riassuntive indicano per gli item un valor medio pari a 0,0 e una *SD* pari a 1,178, per gli individui un valor medio pari a 0,467 e una SD pari a 1,065. Il Fit dei residui è soddisfacente sia per gli item (*mean* = 0,053 e *SD* = 0,989) che per gli individui (*mean* = -0,131 e *SD* = 0,899) nonostante ci sia ancora una leggera asimmetria positiva e una curtosi negativa. L'Indice di Separazione è 0,800 e il *Cronbach Alpha* 0,804 (figura 4.39).

Tutti gli item risultano avere un *fit* e un valore di *p-value* nei limiti (fig. 4.40) e tutte le *ICC* confermano un andamento corretto delle probabilità di risposta osservate; i pochissimi soggetti che presentavano un *Fit Residual* troppo elevato o troppo basso sono stati rimossi dal campione.

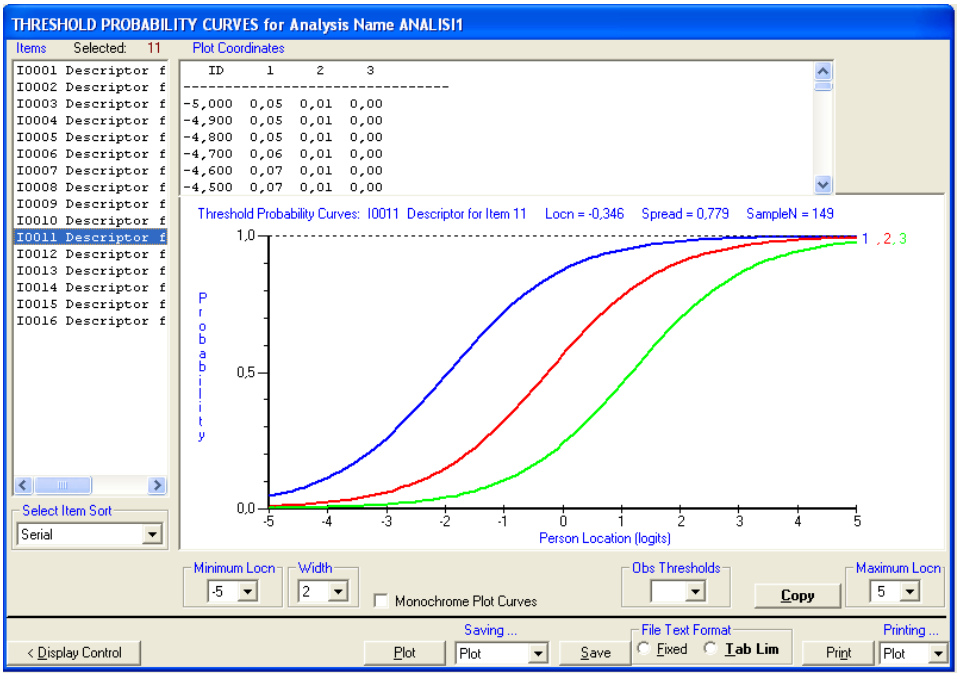


Figura 4.34: Thresholds Probability Curves nel caso politomico

4.3 La costruzione della scala di misura

4.3.1 Legare e ancorare gli item di test diversi

La maggior parte delle analisi quantitative in campo pisico-socio-educativo richiede di poter confrontare nel tempo e nello spazio i livelli di determinate caratteristiche individuali misurate con test o questionari; in ambito educativo, per esempio, si può presentare la necessità di studiare come si sviluppano le conoscenze/competenze/abilità di coorti di studenti nei vari stadi dell'apprendimento scolastico, oppure di confrontare i livelli di gruppi di discenti appartenenti a scuole, a regioni o addirittura a stati diversi. È praticamente impossibile, quando non addirittura concettualmente sbagliato, utilizzare lo stesso questionario per misurare gruppi di studenti di livelli d'apprendimento diversi o lo stesso gruppo di studenti in epoche diverse. Dovendo pertanto ricorrere all'utilizzo di più test per misurare livelli differenti dello stesso tratto latente, sorge il problema di come poter confrontare e interpretare i risultati delle diverse prove.

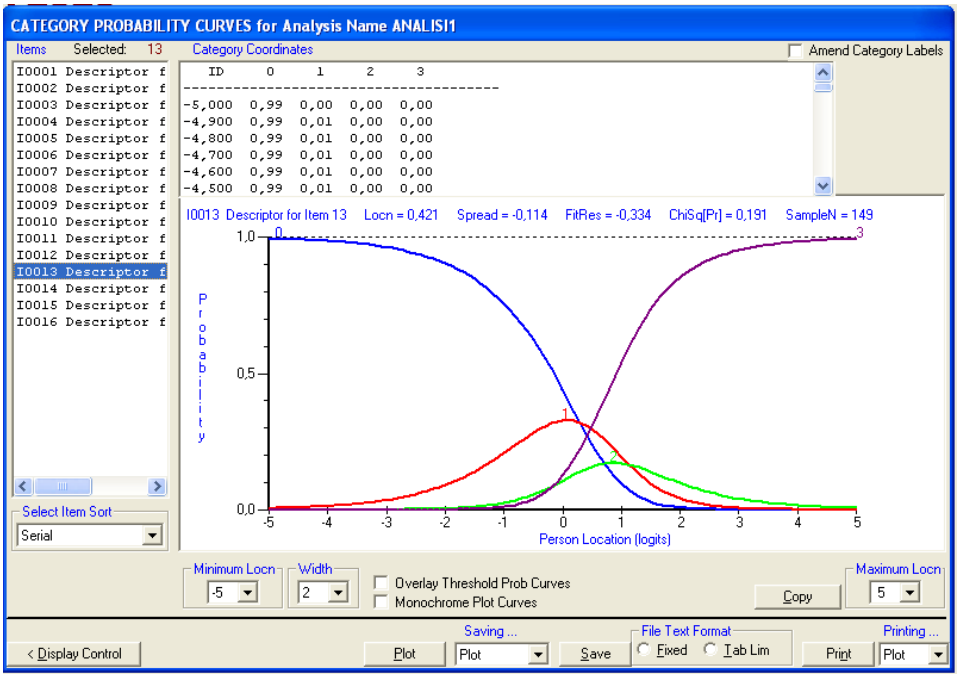


Figura 4.35: Category Probability Curves con Misfit nel caso politomico

Lo sviluppo della Teoria Moderna dei Test con i Modelli di Rasch e il progresso dell'IT hanno permesso di trovare soddisfacenti risposte a tali quesiti. Si è visto che una delle caratteristiche precipue di tutti i RM è l'*oggettività specifica* in base alla quale, all'interno di una stessa struttura di riferimento, il confronto tra le abilità di individui è indipendente dagli item utilizzati e viceversa. Tale requisito di cui gode lo strumento di misura consente di confrontare non solo studenti della stessa coorte a cui viene somministrato lo stesso test ma anche studenti della stessa coorte a cui sono assegnati test diversi o studenti appartenenti a coorti diverse; infatti è possibile stimare i parametri del modello (nella fattispecie le difficoltà degli item di prove diverse) anche in presenza di interi blocchi di dati mancanti da sistema (purché ci sia un sufficiente numero di item di aggancio - *link* - tra le diverse prove). Il fatto poi che sia necessario imporre sempre un'origine arbitraria alla scala del *continuum* ogniqualvolta gli item e gli individui sono misurati e che quindi scale diverse non sono di per sé immediatamente comparabili, viene risolto con le tecniche di ancoraggio che consentono di definire un'origine diversa, fissando le stime delle difficoltà degli

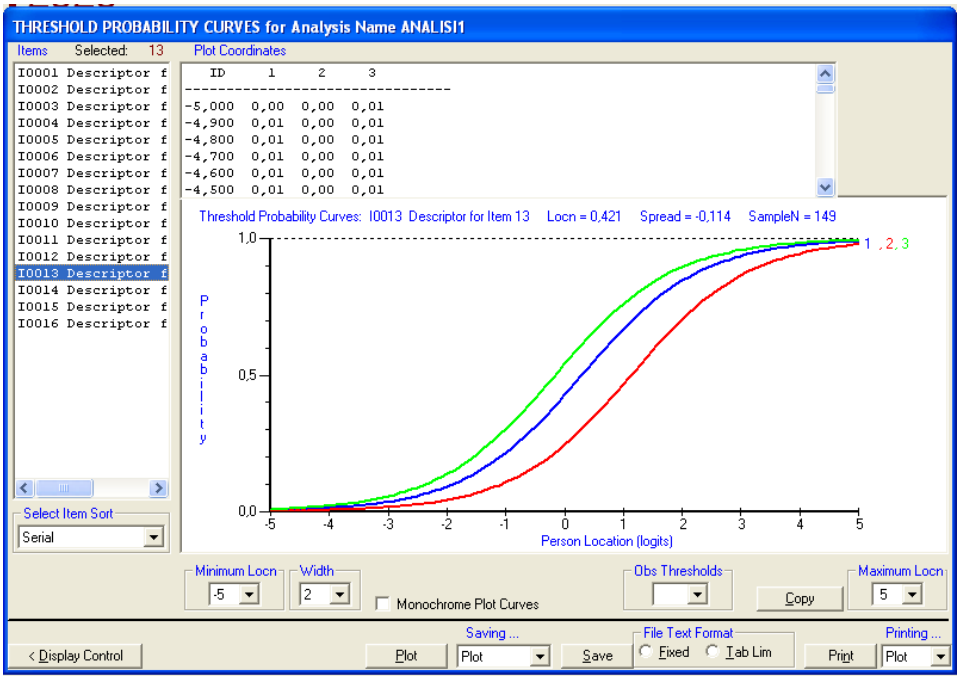


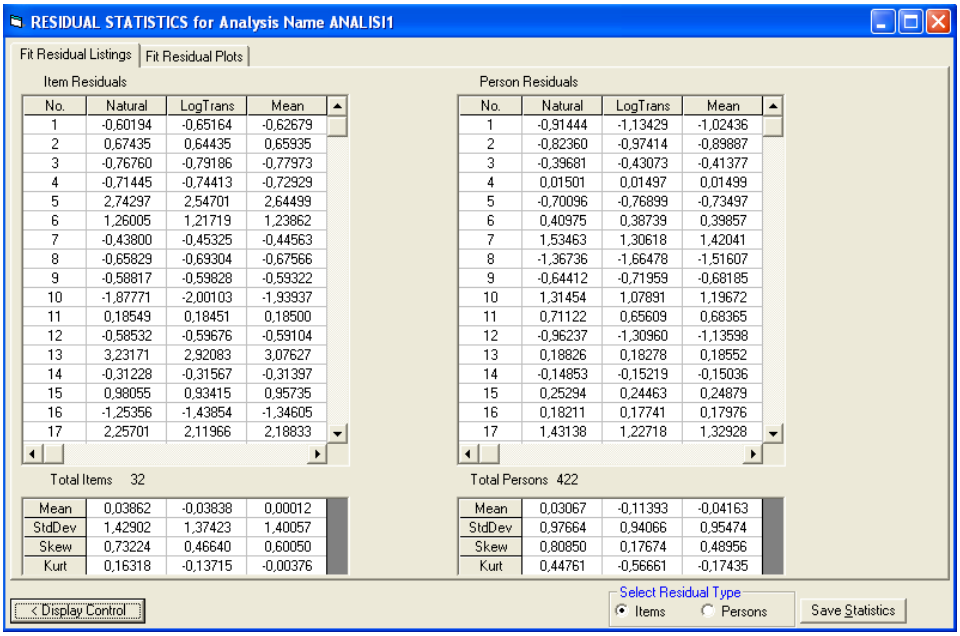
Figura 4.36: Thresholds Probability Curves con Misfit nel caso politomico

item a valori predefiniti (per esempio a misure stimate precedentemente). La tecnica dell'ancoraggio ha la sua applicazione principale nel *test equating*, cioè nel rappresentare sulla medesima scala metrica item di test differenti, quando i dati storici originari non sono disponibili. RUMM2020 e Winsteps hanno implementati degli algoritmi che consentono di legare tra loro test diversi oppure di ancorare determinati item a valori prefissati (Andrich, Sheridan & Luo, 2000; Luo, Seow & Chin, 2001).

In RUMM2020 la procedura di stima dei parametri del modello si basa sull'algoritmo per dati appaiati che segue tre stadi:

1. calcolare le statistiche sufficienti per i parametri degli item;
2. stimare i parametri degli item (posizione, unità, asimmetria e curtosi);
3. stimare le abilità dei soggetti utilizzando i valori di stima degli item.

L'algoritmo di stima per dati appaiati non è compromesso dalla presenza di dati mancanti da sistema, quindi la procedura per legare due o più test è relativamente semplice e risulta sicuramente preferibile ogniqualvolta si hanno



No.	Natural	LogTrans	Mean
1	-0,60194	-0,65164	-0,62679
2	0,67435	0,64435	0,65935
3	-0,76760	-0,79186	-0,77973
4	-0,71445	-0,74413	-0,72929
5	2,74297	2,54701	2,64499
6	1,26005	1,21719	1,23862
7	-0,43800	-0,45325	-0,44563
8	-0,65829	-0,69304	-0,67566
9	-0,58817	-0,59828	-0,59322
10	-1,87771	-2,00103	-1,93937
11	0,18549	0,18451	0,18500
12	-0,58532	-0,59676	-0,59104
13	3,23171	2,92083	3,07627
14	-0,31228	-0,31567	-0,31397
15	0,98055	0,93415	0,95735
16	-1,25356	-1,43854	-1,34605
17	2,25701	2,11966	2,18833

Mean	0,03862	-0,03838	0,00012
StdDev	1,42902	1,37423	1,40057
Skew	0,73224	0,46640	0,60050
Kurt	0,16318	-0,13715	-0,00376

No.	Natural	LogTrans	Mean
1	-0,91444	-1,13429	-1,02436
2	-0,82360	-0,97414	-0,89887
3	-0,39681	-0,43073	-0,41377
4	0,01501	0,01497	0,01499
5	-0,70096	-0,76899	-0,73497
6	0,40975	0,38739	0,39857
7	1,53463	1,30618	1,42041
8	-1,36736	-1,66478	-1,51607
9	-0,64412	-0,71959	-0,68185
10	1,31454	1,07891	1,19672
11	0,71122	0,65609	0,68365
12	-0,96237	-1,30960	-1,13598
13	0,18826	0,18278	0,18552
14	-0,14853	-0,15219	-0,15036
15	0,25294	0,24463	0,24879
16	0,18211	0,17741	0,17976
17	1,43138	1,22718	1,32928

Mean	0,03067	-0,11393	-0,04163
StdDev	0,97664	0,94066	0,95474
Skew	0,80850	0,17674	0,48956
Kurt	0,44761	-0,56661	-0,17435

Figura 4.37: Statistiche dei residui

a disposizione i risultati completi delle prove. In questo caso si crea un'unica matrice che contiene tutti i dati dei test da "legare"; in ogni casella della matrice, individuata da un item e da un individuo, compare la risposta osservata se l'item appartiene al test che è stato somministrato all'individuo, un valore mancante altrimenti. La matrice viene passata al programma che procede a stimare tutti i parametri degli item e delle persone con un'unica analisi. La costruzione di sistemi di collegamento tra test può essere effettuata in tre modi: mediante una struttura a catena (4.43), un loop (4.44), una rete di loop o test di collegamento intermedi per test paralleli. Nel caso oggetto della presente ricerca si è creato un test di aggancio per legare il test del SNV 2004/2005 al test del SNV 2005/2006 mediante una struttura a catena.

Qualora non si disponga dei dati storici completi dei test da collegare ma solamente delle stime dei parametri dei loro item (posizione, unità, asimmetria e curtosi oppure posizione e soglie) si ricorre alla tecnica dell'ancoraggio. Supponiamo che il Test A sia il test somministrato in passato di cui si hanno solo le stime dei parametri degli item, che il Test B sia la prova attuale di cui si dispongono i risultati e che si vogliano confrontare le abilità dei soggetti

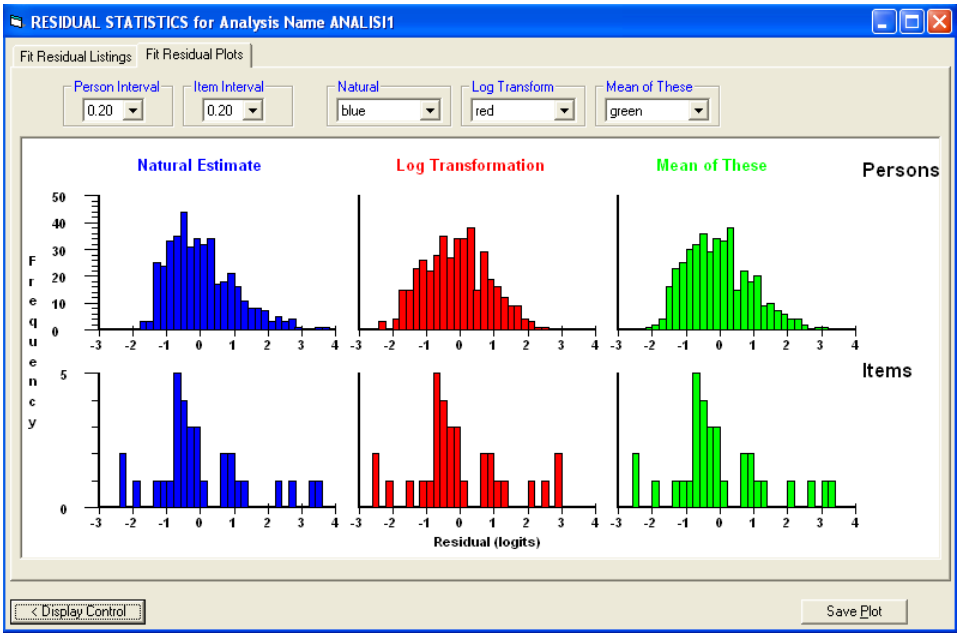


Figura 4.38: Distribuzioni dei residui

misurate nelle due prove. Poiché il Test A e il Test B sono stati analizzati separatamente, le due scale del *continuum* prodotte non coincidono, avendo ciascuna uno zero arbitrariamente fissato sulla media delle difficoltà stimate dei propri item; in questo caso è necessario ancorare l'origine di una scala all'origine dell'altra (per esempio quella del Test B a quella del Test A). La procedura, che va sotto il nome di Analisi dell'Ancoraggio Relativo (*Relative Anchoring Analysis*), consiste nel:

1. identificare i K item comuni dei due test;
2. considerare le loro difficoltà stimate nel Test A;
3. calcolare la loro media M_A;
4. stimare tutte le difficoltà degli item del Test B;
5. calcolare M_B, media delle stime delle difficoltà degli item comuni nel Test B, e $M = M_B - M_A$;
6. adeguare l'origine della scala costruita con il Test B traslando tutte le stime degli item di M.

ITEM-PERSON INTERACTION

	ITEMS			PERSONS		
	Location	Fit	Residual	Location	Fit	Residual
Mean	0,000	0,053		0,467	-0,131	
SD	1,178	0,989		1,065	0,899	
Skewness		0,286			0,204	
Kurtosis		-0,256			-0,471	
Correlation		0,093			-0,158	
Complete data DF =		0,962				

ITEM-TRAIT INTERACTION		RELIABILITY INDICES	
Total Item Chi Squ	94,457	Separation Index	0,80028
Total Deg of Freedom	84,000	Cronbach Alpha	0,80409
Total Chi Squ Prob	0,204164		

Figura 4.39: Summary Statistics del Test Pulito

7. misurare le abilità degli individui utilizzando le stime delle difficoltà trovate al punto 6.

La difficoltà stimata del generico item i_B adeguata alla nuova scala diventa:

$$\widetilde{\delta}_{i_B} = \delta_{i_B} - M. \tag{4.8}$$

Dopo l'adeguamento la media delle difficoltà degli item del Test B diventa:

$$\frac{\sum_{i_B=1}^{I_B} \widetilde{\delta}_{i_B}}{I_B} = \frac{\sum_{i_B=1}^{I_B} (\delta_{i_B} - M)}{I_B} = \frac{\sum_{i_B=1}^{I_B} \delta_{i_B} - I_B M}{I_B} = \frac{0 - I_B M}{I_B} = -M \tag{4.9}$$

dove I_B è il numero di item del Test B. Sempre nel Test B la media delle stime delle difficoltà degli item comuni diventa:

$$\frac{\sum_{i_B=1}^{K} \widetilde{\delta}_{i_B}}{K} = \frac{\sum_{i_B=1}^{K} (\delta_{i_B} - M)}{K} = \frac{\sum_{i_B=1}^{K} \delta_{i_B} - KM}{K} = \frac{KM_B - K(M_B - M_A)}{K} = M_A \tag{4.10}$$

RUMM2020 Project: Test 2008/2009 Analysis: ANALISISI
Title: SENZA 5 10 13 15 18 30
Display: INDIVIDUAL ITEM-FIT - Serial Order

Seq	Item	Type	Location	SE	Residual	DF	ChiSq	DF	Prob
1	d01	Poly	-2,163	0,179	-0,230	403,85	2,660	3	0,447016
2	d02	Poly	-1,166	0,132	0,545	403,85	3,482	3	0,323047
3	d03	Poly	-0,218	0,112	-0,087	403,85	3,556	3	0,313539
4	d04	Poly	-0,761	0,121	-0,599	403,85	6,319	3	0,097099
6	d06	Poly	0,636	0,108	2,004	403,85	8,457	3	0,037460
7	d07	Poly	-1,294	0,136	-0,580	403,85	1,454	3	0,693036
8	d08	Poly	-1,277	0,136	-0,865	403,85	2,618	3	0,454399
9	d09	Poly	0,477	0,108	-0,805	403,85	0,661	3	0,882313
11	d11	Poly	0,761	0,108	0,105	403,85	2,248	3	0,522543
12	d12	Poly	0,110	0,109	-0,296	403,85	3,835	3	0,279848
14	d14	Poly	0,011	0,109	-0,182	403,85	2,838	3	0,417308
16	d16	Poly	-1,737	0,155	-1,304	403,85	8,071	3	0,044570
17	d17	Poly	0,852	0,109	2,439	403,85	2,572	3	0,462479
19	d19	Poly	0,003	0,109	-0,959	403,85	1,659	3	0,646140
20	d20	Poly	0,531	0,108	-0,566	403,85	1,425	3	0,699755
21	d21	Poly	0,641	0,108	-2,167	403,85	7,998	3	0,046051
22	d22	Poly	2,302	0,136	-0,056	403,85	0,978	3	0,806514
23	d23	Poly	0,944	0,109	0,960	403,85	0,262	3	0,966951
24	d24	Poly	-0,179	0,111	0,375	403,85	6,209	3	0,101888
25	d25	Poly	0,706	0,108	0,536	403,85	2,920	3	0,404146
26	d26	Poly	-1,052	0,128	0,112	403,85	1,245	3	0,742336
27	d27	Poly	-0,310	0,113	0,570	403,85	1,472	3	0,688762
28	d28	Poly	1,550	0,117	-0,232	403,85	3,904	3	0,272060
29	d29	Poly	-1,106	0,130	-0,581	403,85	3,367	3	0,338413
31	d31	Poly	2,342	0,137	1,208	403,85	5,823	3	0,120563
32	d32	Poly	-0,602	0,118	0,556	403,85	5,573	3	0,134363

Figura 4.40: Individual Item Fit test 2008/2009 Pulito

La procedura dà buoni risultati quando sia il Test A che il Test B si adattano bene al RM; è però sufficiente che uno dei due test ne violi alcuni requisiti perché l'ancoraggio produca misure distorte.

Un'altra tecnica di ancoraggio, utilizzata quando non si hanno a disposizione i dati originali completi e si vogliono mantenere inalterate le stime storiche degli item di aggancio, va sotto il nome di Analisi dell'Ancoraggio Assoluto (*Absolute Anchoring Analysis*). La procedura consiste nel:

1. identificare i K item comuni dei due test;

2. prendere le difficoltà stimate nel Test A degli item comuni;

Items / Persons	1 60	61 120
1	**Test A**	**Test B**
person		
1200		

Figura 4.41: Due test della stessa dimensione (60 item) somministrati a tutti

	Items Gruppo 1	Items Gruppo 2	Items Gruppo 3
Individui Gruppo 1	**Test A**		
Individui Gruppo 2		**Test C**	
Individui Gruppo 3			**Test B**

Figura 4.42: Tre test somministrati a tre gruppi di persone

3. calcolare le statistiche sufficienti per i parametri degli item usando le osservazioni del Test B;

4. stimare tutti i parametri del test B, ma ad ogni ciclo della procedura di stima sostituire le stime degli item comuni con i valori trovati nel Test A, finché non si raggiunge la convergenza;

5. stimare le abilità degli individui usando i valori trovati.

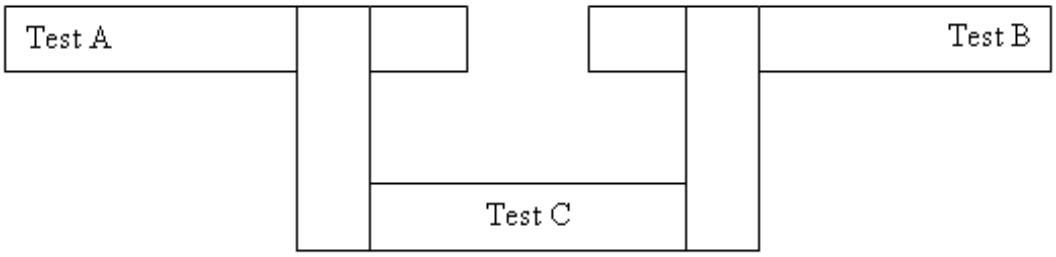


Figura 4.43: Una catena con due agganci

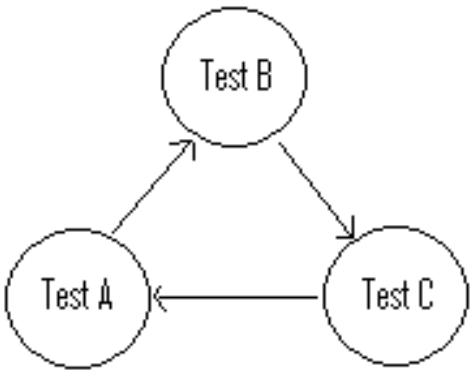


Figura 4.44: Un loop di tre test

4.3.2 L'unità di misura nel Modello di Rasch

Nelle scienze fisiche "a measurement of a magnitude of a quantitative attribute is an estimate of the ratio between that magnitude and whichever magnitude of the same attribute is taken as the unit of measurement" (Michell, 1997). L'unità di misura assume un ruolo centrale nella misurazione: essa determina il fattore di separazione tra i valori numerici delle misurazioni prese con uno strumento di misura. Nel RM l'unità di misura riveste un ruolo che rimane implicito, se il tratto latente da misurare viene considerato in un'unica struttura di riferimento ma che si esplicita se lo stesso costrutto viene considerato all'interno di due o più strutture di riferimento. Rasch ha definito una struttura di riferimento (*frame of reference*) come "a class of persons respon-

ding to a class of item in a well-defined assessment context" (Rasch, 1977). All'interno di una struttura di riferimento la misurazione oggettiva richiede che venga soddisfatta l'invarianza dei confronti: "The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared .... Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion" (Rasch, 1961). Si è già osservato che l'espressione matematica del RM è necessaria e sufficiente a garantire tale condizione (Andersen, 1977).

Consideriamo ora il ruolo dell'unità di misura nel RM da un punto di vista algebrico. Nel caso dicotomico il RM è specificato dalla:

$$Pr\{X_{\nu i} = 1\} = \frac{exp(\theta_\nu - \delta_i)}{1 + exp(\theta_\nu - \delta_i)} \tag{4.11}$$

Se si pone lo *score* dell'individuo ν $r_\nu = x_{\nu i} + x_{\nu j}$ risulta che:

$$Pr\{(X_{\nu i} = 1; X_{\nu j} = 0)|r_\nu = 1\} = \frac{exp(\delta_j - \delta_i)}{1 + exp(\delta_j - \delta_i)} \tag{4.12}$$

e il suo complemento

$$Pr\{(X_{\nu i} = 0; X_{\nu j} = 1)|r_\nu = 1\} = \frac{1}{1 + exp(\delta_j - \delta_i)} \tag{4.13}$$

Le probabilità della 4.11 e della 4.12 non dipendono da θ_ν, cioè il punteggio totale r_ν è una statistica sufficiente θ_ν. Dalle due si ricava:

$$\ln\left(\frac{Pr\{(X_{\nu i} = 1, X_{\nu j} = 0|r_\nu = 1)\}}{Pr\{(X_{\nu i} = 0, X_{\nu j} = 1|r_\nu = 1)\}}\right) = \delta_j - \delta_i \tag{4.14}$$

che esprime la differenza dei parametri degli item, anch'essa indipendente dai parametri delle persone.
Indicando con F_{ij} la frequenza osservata del *pattern* di risposta ($X_{\nu i} = 1, X_{\nu j} =$

0), con F_{ji} la frequenza osservata del *pattern* di risposta ($X_{\nu i} = 0, X_{\nu j} = 1$) e con $\delta_{ji} = \delta_j - \delta_i$ dalla equazione

$$ln(\frac{F_{ij}}{F_{ji}}) = \widehat{\delta}_{ji} \tag{4.15}$$

si ottiene una stima della differenza δ_{ji}, generalmente indicata come δ_{ji} *logit*. L'unità implicita nella stima di $\widehat{\delta}_{ji}$ è determinata da una scelta arbitraria della costante moltiplicativa nell'equazione del modello. Rendendo esplicita tale costante la 4.11 si può riscrivere come:

$$Pr\{X_{\nu i} = 1\} = \frac{exp\{\rho(\theta_\nu/\rho - \delta_i/\rho)\}}{1 + exp\{\rho(\theta_\nu/\rho - \delta_i/\rho)\}} \tag{4.16}$$

$$= \frac{exp\{\rho(\theta_{*\nu} - \delta_{*i})\}}{1 + exp\{\rho(\theta_{*\nu} - \delta_{*i})\}} \tag{4.17}$$

dove $\beta_{*\nu} = \beta_\nu/\rho$, $\delta_{*i} = \delta_i/\rho$ e ρ è la costante moltiplicativa.
Se si considera un'unica struttura di riferimento il valore di ρ è necessariamente arbitrario; solitamente si pone pari a 1. L'espressione della differenza di parametri dei due item diventa:

$$ln(\frac{F_{ij}}{F_{ji}}) = \rho(\widehat{\delta}_{*ji})$$

dove $\widehat{\delta}_{*ji} \equiv \widehat{\delta}_{*j} - \widehat{\delta}_{*i}$. Tale stima è ancora una volta indipendente dai parametri delle persone, e poiché la parte a sinistra dell'equazione è un numero reale e la parte a destra ha due fattori, è evidente anche per la stima $\widehat{\delta}_{*ji}$ la scelta dell'unità della scala è arbitraria.

Volendo misurare lo stesso tratto latente all'interno di due strutture di riferimento diverse, è possibile, e spesso necessario, operare dei confronti tra le unità di misura (rimanendo valida la proprietà della sufficienza del modello): bisogna conoscere la relazione tra le unità di misura al fine di convertire una misurazione effettuata con uno strumento nell'unità di un altro strumento. Nelle scienze fisiche la relazione tra unità di misura è nota a priori, viceversa nel RM tale relazione, generalmente, non è nota ma si può determinare applicando lo stesso principio che si utilizza nelle scienze fisiche: il confronto tra le unità

di misura si realizzare attraverso il confronto tra le misurazioni dello stesso oggetto. Il rapporto tra le misurazioni, infatti, è inversamente proporzionale al rapporto delle grandezze delle unità implicite nelle misurazioni (Humphry, Andrich, 2007; Humphry, 2005) ed è perciò possibile determinare il fattore di scala $\rho_{S_1 \to S_2}$ per convertire le misurazioni effettuate in due unità differenti (S_1 e S_2), considerando:

$$\rho_{S_1 \to S_2} = \frac{m_{S_2}}{m_{S_1}}$$

dove m è l'oggetto misurato nei due sistemi. In maniera analoga si ha:

$$m_{S_2} = \rho_{S_1 \to S_2} m_{S_1}.$$

Nel RM la grandezza che si vuole misurare è la differenza tra le difficoltà di due item, cioè d_{ji} l'intervallo tra l'item j e l'item i sul *continuum* latente, invariante dalla struttura di riferimento. Una misura di d_{ji} è data dal rapporto di d_{ji} con qualsiasi intervallo preso come unità di misura. Tale rapporto è un numero reale. Prese due diverse unità di misura (dello stesso costrutto) u_{S1} e u_{S2} che rappresentano due diverse grandezze mediante le quali il *continuum* viene partizionato, si ha:

$$\delta_{jiS_1} = \frac{d_{ji}}{u_{S_1}}$$

e

$$\delta_{jiS_2} = \frac{d_{ji}}{u_{S_2}}$$

che sono le due misure della stessa grandezza rappresentate sulla stessa scala di misura (retta dei numeri reali) partizionata in base a due sistemi diversi, in cui l'origine è fissata arbitrariamente mentre il fattore moltiplicativo di separazione è determinato empiricamente dall'unità di misura stessa, e

$$d_{ji} = \delta_{jiS_1} u_{S_1} = \delta_{jiS_2} u_{S_2}.$$

Nel RM a ogni sistema di riferimento corrisponde una specifica unità di misura, implicitamente determinata dal contesto in cui avviene la misurazione cioè dall'osservazione dei dati delle risposte risultato dall'interazione delle persone con gli item. Tale unità di misura costituisce il fattore di separazione

tra le misurazioni che sono espresse in valori *logit*.

Nonostante l'indubbio interesse teorico della questione, a fini pratici è possibile risolvere il problema ristimando abilità e difficoltà sul *data set* comprendente tutti gli item e tutti gli individui.

4.3.3 Il link dei tre test

La costruzione della scala di misurazione è stata effettuata analizzando congiuntamente i test del SNV 2004/2005, del SNV 2005/2006 e il test di *link* in modo da ottenere le stime di tutti gli item sulla stessa scala di misura e con la stessa unità di misura. A tale scopo sono stati selezionati due campioni di 400 studenti ciascuno dai database del SNV puliti dai casi di *cheating* e di *misfit* individuali; uno dalla popolazione del 2004/2005 senza i tre item (9, 11 e 20) che presentavano *misfit* e uno da quella del 2005/2006 anche in questo caso dopo aver eliminato i 3 item (6, 10 e 16) che mostravano cattivo adattamento al modello. A questi due campioni sono state aggiunte le osservazioni raccolte con la somministrazione del test di link eliminando i 6 item (5, 10, 13, 15, 18 e 30) che hanno evidenziato malfunzionamento e i 22 soggetti che meno si adattavano al modello.

La figura 4.46 rappresenta le *Item-Person Map* dei tre test (il test di link contiene tutti e 32 gli item originari) allineate sulle rispettive origini delle scale. La figura 4.47, invece, rappresenta la *Item-Person Map* finale costruita con i 64 item rimanenti; gli item a01, ..., a28 (senza gli item eliminati ed evidenziati in colore azzurro) sono gli item del SNV 2004/2005, gli item b01, ..., b28 (senza gli item eliminati ed evidenziati in colore fucsia) sono gli item del SNV 2005/2006 e gli item link01, ..., link26 (senza gli item eliminati) sono quelli relativi al test di link. La tabella 4.3, infine, rappresenta le misure di tutti i 64 item della scala calcolate da RUMM2020; nell'ordine contiene: il nome dell'item (*Item*), la difficoltà stimata (*Location*), il suo *standard error* (*SE*), il suo residuo standardizzato (*Residual*) coi sui gdl (*DF*) e la statistica chi-quadrato con i suoi gdl e il suo *p-value* (*Prob*). La figura 4.5 rappresenta le stesse misure calcolate da Winsteps; nell'ordine compaiono: il nome dell'item (*NAME*), la difficoltà stimata (*MEASURE*), l'errore di misura (*ERROR*),

l'*Infit Mean Square* (*IN.MSQ*) e il suo valore standardizzato e l'*Outfit Mean Square* (*OUT.ZST*) e il suo valore standardizzato.

4.3.4 Le analisi sulla bontà della scala

Sicuramente il test per costruire la scala è risultato troppo facile per valutare con precisione le abilità più elevate; si vede infatti dalla figura 4.47 che quasi un logit separa la media delle stime degli item dalla media delle stime degli studenti. Sicuramente questa anomalia è dovuta al fatto che, come si è visto, sia la prova del 2004/2005 sia quella del 2005/2006 sono risultate troppo facili. Inoltre però, ha inciso il fatto che i 6 item del test di link eliminati per il loro cattivo adattamento erano tutti di difficoltà medio-alta; la loro rimozione ha determinato un ulteriore slittamento verso il basso di tutto il set di item rimanente dopo tali esclusioni.

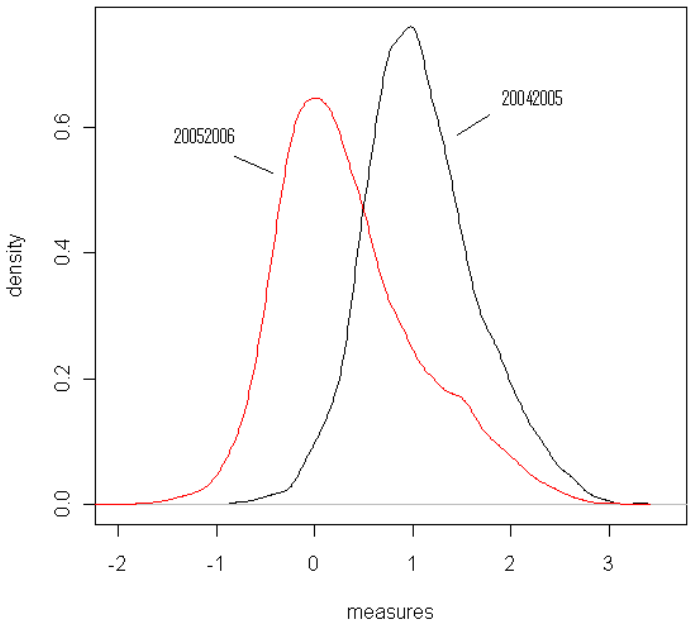


Figura 4.45: Curve di densità delle abilità medie scolastiche

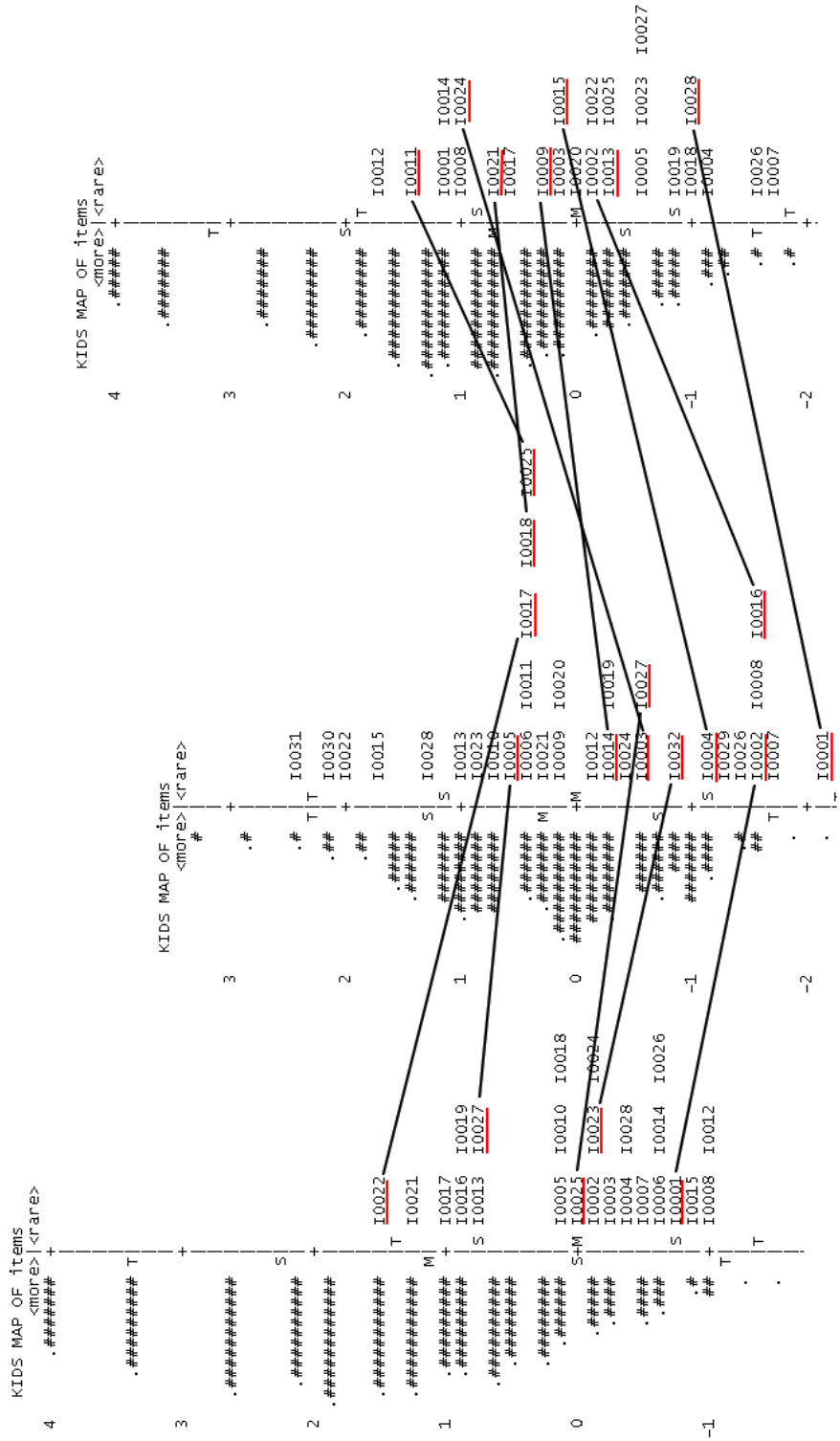


Figura 4.46: I tre test con gli item di aggancio

4.4 Un'applicazione della scala di misura

Come emerge dai rapporti annuali sui risultati della rilevazione degli apprendimenti degli studenti, l'INVALSI svolge due tipi di analisi:

- una a livello "macro" in cui descrive le performance degli studenti a livello di macro aree, regioni e province, confrontando le medie (le percentuali di risposta di gruppi di item) degli *score* grezzi o standardizzati tra di loro o con i livelli medi regionali,
- l'altra a livello "micro", indicando che posizione occupa ogni scuola nella distribuzione degli *score* medi (in sostanza tra quali percentili si colloca), a livello nazionale, regionale e di macro area.

Dichiaratamente si esime dal fornire alcun risultato fondato su analisi di tipo diacronico, non essendo in grado di confrontare i dati dei vari test in quanto le prove non sono costituite da item "equivalenti", e di conseguenza non dà alcuna indicazione nè sui livelli di apprendimento raggiunti alle diverse epoche nè sulle variazioni delle conoscenze/competenze (il valore aggiunto) tra un anno e l'altro. In aggiunta, poiché, come si è visto, molti dati sono affetti da *cheating*, anche i risultati basati sugli *score* grezzi risultano poco attendibili. Lo studio condotto sfruttando la metodologia dell'analisi di Rasch ha permesso di agganciare tra loro i test dei vari anni e di confrontare, nello spazio e nel tempo, le abilità in matematica degli studenti del IV anno della scuola primaria. Qui di seguito si espongono alcuni risultati applicativi delle analisi svolte utilizzando le misure stimate. Bisogna comunque sottolineare il fatto che la procedura di pulizia dei dati ha migliorato, in parte, l'affidabilità delle informazioni deducibili dai questionari del Sistema Nazionale di Valutazione, ma non ha risolto tutti i problemi che stavano alla base. Inoltre le prove del SNV considerate sono state somministrate in mesi differenti (in aprile 2005 quelle del 2004/2005 e a cavallo tra novembre e dicembre 2005 quelle del 2005/2006): quindi il confronto tra gli apprendimenti dei due anni risulta falsato perché si riferisce a due momenti del processo di apprendimento diversi. Ciò non di meno si ritiene che l'approccio utilizzato e i risultati ottenuti forniscano una valida traccia che l'INVALSI (o un qualsiasi altro ente di valutazione del sistema

Item	Location	SE	Residual	DF	ChiSq	DF	Prob
b26	-2,227	0,165	0,309	375,26	0,582	3	0,9006
b07	-2,023	0,155	0,39	375,26	0,741	3	0,8635
b28 - link01	-1,987	0,124	-1,265	757,23	3,712	3	0,2942
b04	-1,798	0,146	0,002	375,26	10,923	3	0,0121
b18	-1,688	0,142	-1,187	375,26	1,759	3	0,6239
b19	-1,378	0,133	-0,625	375,26	1,842	3	0,6058
b27	-1,154	0,127	-0,669	375,26	1,713	3	0,6339
b23	-1,052	0,125	0,428	375,26	4,036	3	0,2577
b13 - link16	-1,039	0,096	-1,212	757,23	3,281	3	0,3503
a08	-1,008	0,154	0,946	361,86	6,309	3	0,0975
b05	-0,972	0,123	0,216	375,26	0,566	3	0,9042
a12	-0,959	0,152	0,644	361,86	6,169	3	0,1037
b22	-0,867	0,121	0,326	375,26	2,22	3	0,5280
link07 (d3 1m 0405)	-0,825	0,142	-0,553	381,96	1,519	3	0,6779
a01 - link02	-0,823	0,102	0,432	743,82	6,398	3	0,0938
link08 (d27 1m 0506)	-0,78	0,14	-0,869	381,96	3,498	3	0,3211
b25	-0,765	0,12	0,217	375,26	1,836	3	0,6071
a15	-0,741	0,144	-0,176	361,86	2,433	3	0,4876
b02	-0,724	0,119	0,431	375,26	1,441	3	0,6960
a07	-0,676	0,141	-0,257	361,86	0,489	3	0,9213
link29 (d30 1m 0506)	-0,631	0,135	-0,455	381,96	3,05	3	0,3840
link26 (TIMSS)	-0,605	0,134	0,323	381,96	0,677	3	0,8786
a06	-0,463	0,135	-1,68	361,86	2,952	3	0,3991
a26	-0,448	0,135	0,642	361,86	1,818	3	0,6111
a14	-0,405	0,134	0,48	361,86	1,269	3	0,7364
a04	-0,396	0,133	-2,224	361,86	9,915	3	0,0193
b20	-0,348	0,115	1,876	375,26	8,189	3	0,0423
b15 - link04	-0,302	0,085	-0,15	757,23	1,942	3	0,5846
b03	-0,293	0,115	-0,019	375,26	2,137	3	0,5445
b17	-0,192	0,114	1,458	375,26	3,288	3	0,3493
a24	-0,175	0,128	-1,084	361,86	6,986	3	0,0723
a03	-0,112	0,127	-1,933	361,86	10,819	3	0,0127

Tabella 4.3: Gli item della scala di misura stimate da RUMM2020 - parte 1

Item	Location	SE	Residual	DF	ChiSq	DF	Prob
a02	-0,083	0,127	1,506	361,86	4,553	3	0,2076
a28	-0,075	0,126	0,944	361,86	2,559	3	0,4647
a23 - link32	0,083	0,085	1,656	743,82	4,191	3	0,2415
a25 - link27	0,107	0,085	0,65	743,82	5,651	3	0,1299
a05	0,162	0,123	0,696	361,86	1,254	3	0,7401
a10	0,199	0,122	-0,158	361,86	0,672	3	0,8799
b09 - link14	0,258	0,081	1,399	757,23	2,724	3	0,4361
a18	0,301	0,121	-0,095	361,86	0,666	3	0,8813
link24 (d13 1m 0506)	0,306	0,115	0,387	381,96	7,268	3	0,0638
b08	0,352	0,113	0,191	375,26	3,354	3	0,3402
b01	0,41	0,113	0,006	375,26	2,24	3	0,5242
b24 - link03	0,478	0,08	-0,697	757,23	6,506	3	0,0894
link19 (TIMSS)	0,484	0,113	-0,672	381,96	1,599	3	0,6596
link12 (TIMSS)	0,577	0,112	-0,043	381,96	4,363	3	0,2248
b14	0,643	0,115	-1,663	375,26	6,439	3	0,0921
a13	0,839	0,117	-2,509	361,86	14,898	3	0,0019
a19	0,898	0,117	0,141	361,86	0,514	3	0,9159
link09 (TIMSS)	0,931	0,111	-0,487	381,96	0,561	3	0,9053
a16	0,961	0,117	-0,797	361,86	3,033	3	0,3866
link20 (TIMSS)	1,012	0,111	-0,38	381,96	0,853	3	0,8367
a17	1,051	0,117	-2,126	361,86	6,196	3	0,1024
b11 - link25	1,063	0,081	0,255	757,23	6,414	3	0,0931
link21 (scarto V)	1,107	0,111	-1,892	381,96	5,793	3	0,1221
link06 (d3 1m 0506)	1,15	0,111	2,019	381,96	8,514	3	0,0365
link11 (TIMSS)	1,232	0,111	0,328	381,96	2,165	3	0,5389
a21	1,254	0,117	-1,553	361,86	11,018	3	0,0116
b12	1,3	0,123	0,671	375,26	2,776	3	0,4274
a22 - link17	1,404	0,081	2,407	743,82	2,392	3	0,4952
link23 (scarto V)	1,473	0,113	0,854	381,96	1,186	3	0,7563
link28 (TIMSS)	2,077	0,121	-0,473	381,96	2,997	3	0,3920
link22 (TIMSS)	2,824	0,14	-0,164	381,96	1,039	3	0,7919
link31 (scarto V)	3,08	0,149	-0,034	381,96	2,702	3	0,4399

Tabella 4.4: Gli item della scala di misura stimate da RUMM2020 - parte 2

NAME	MEASURE	ERROR	IN.MSQ	IN.ZST	OUT.MSQ	OUT.ZST
b26	-2,20	0,17	1,02	0,21	1,16	0,72
b07	-2,01	0,16	1,08	0,75	1,17	0,84
b28 - link01	-1,89	0,12	0,93	-0,88	0,76	-1,43
b04	-1,70	0,15	0,93	-0,76	1,02	0,20
b18	-1,60	0,14	0,95	-0,52	0,81	-1,14
b19	-1,29	0,13	0,95	-0,69	0,92	-0,56
a08	-1,24	0,16	1,08	0,81	1,46	1,84
a12	-1,19	0,16	1,08	0,87	1,29	1,25
b27	-1,06	0,13	0,92	-1,33	0,93	-0,57
b13 - link16	-0,98	0,10	0,96	-0,77	0,87	-1,11
b23	-0,97	0,13	0,94	-0,92	1,10	0,86
a15	-0,94	0,15	1,06	0,68	1,03	0,23
a01 - link02	-0,92	0,10	0,99	-0,14	1,15	1,09
link07 (d3 1m 0405)	-0,88	0,15	1,04	0,49	0,93	-0,32
b05	-0,87	0,12	1,02	0,29	1,02	0,18
a07	-0,83	0,14	0,97	-0,41	1,03	0,23
link08 (d27 1m 0506)	-0,79	0,14	0,96	-0,44	0,86	-0,73
b22	-0,76	0,12	1,00	-0,06	1,06	0,57
b25	-0,67	0,12	1,02	0,36	1,02	0,21
link29 (d30 1m 0506)	-0,66	0,14	0,95	-0,64	0,95	-0,24
link26 (TIMSS)	-0,64	0,14	0,99	-0,15	1,11	0,66
b02	-0,62	0,12	1,03	0,62	1,03	0,38
a26	-0,61	0,14	1,00	0,04	1,24	1,44
a06	-0,59	0,14	0,92	-1,20	0,78	-1,45
a14	-0,57	0,14	1,05	0,70	1,18	1,11
a04	-0,52	0,14	0,89	-1,71	0,72	-2,00
a24	-0,31	0,13	0,92	-1,37	0,90	-0,70
a02	-0,24	0,13	1,10	1,61	1,31	2,17
b15 - link04	-0,24	0,09	0,98	-0,53	0,97	-0,33
b20	-0,24	0,12	1,13	2,55	1,15	1,80
a03	-0,22	0,13	0,88	-2,03	0,79	-1,72
a28	-0,21	0,13	1,01	0,16	1,22	1,67

Tabella 4.5: Gli item della scala di misura stimate da Winsteps - parte 1

NAME	MEASURE	ERROR	IN.MSQ	IN.ZST	OUT.MSQ	OUT.Z
b03	-0,17	0,12	1,01	0,31	0,98	-0,28
b17	-0,08	0,12	1,07	1,45	1,11	1,42
a23 - link32	0,01	0,09	1,08	2,01	1,17	2,06
a05	0,03	0,12	1,06	1,02	1,12	1,08
a25 - link27	0,06	0,09	1,06	1,51	1,08	1,02
a10	0,08	0,12	1,01	0,26	1,00	0,07
a18	0,19	0,12	1,04	0,80	0,99	-0,04
link24 (d13 1m 0506)	0,31	0,12	1,02	0,47	1,03	0,32
b09 - link14	0,34	0,08	1,04	1,16	1,05	0,83
b08	0,50	0,11	0,98	-0,40	1,00	-0,02
link19 (TIMSS)	0,51	0,11	0,98	-0,45	0,92	-0,86
b24 - link03	0,56	0,08	0,93	-2,10	0,97	-0,61
b01	0,58	0,11	1,01	0,13	0,98	-0,23
link12 (TIMSS)	0,61	0,11	1,02	0,58	0,97	-0,32
a13	0,74	0,12	0,90	-2,01	0,81	-2,55
b14	0,76	0,12	0,92	-1,70	0,85	-2,03
a19	0,82	0,12	1,02	0,45	1,02	0,28
a16	0,87	0,12	0,95	-1,04	0,95	-0,61
a17	0,95	0,12	0,91	-1,81	0,85	-2,12
link09 (TIMSS)	0,95	0,11	0,99	-0,27	0,94	-0,81
link20 (TIMSS)	1,03	0,11	0,98	-0,53	0,95	-0,65
b11 - link25	1,12	0,08	0,98	-0,59	0,99	-0,15
a21	1,14	0,12	0,89	-2,29	0,88	-1,65
link21 (scarto V)	1,14	0,11	0,93	-1,79	0,85	-2,18
link06 (d3 1m 0506)	1,23	0,11	1,12	2,77	1,12	1,74
link11 (TIMSS)	1,28	0,11	1,00	-0,06	1,00	0,04
a22 - link17	1,45	0,08	1,07	1,93	1,14	2,60
link23 (scarto V)	1,53	0,11	1,00	0,09	1,05	0,67
b12	1,54	0,13	1,04	0,63	1,11	0,98
link28 (TIMSS)	2,15	0,12	0,96	-0,72	0,94	-0,64
link22 (TIMSS)	2,96	0,14	0,98	-0,15	1,00	0,03
link31 (scarto V)	3,29	0,16	1,08	0,80	1,08	0,48

Tabella 4.6: Gli item della scala di misura stimate da Winsteps - parte 2

educativo) potrebbe seguire se volesse monitorare correttamente, nel tempo e nello spazio, i livelli di apprendimento della matematica, o di qualsiasi altra disciplina, e costituiscano una buona base di partenza per ulteriori sviluppi di ricerca metodologica ed empirica.

4.4.1 Valutazione delle abilità nei due anni

Dall'analisi del test di *link* è risultato che le difficoltà medie dei due test erano diverse; il test 2004/2005 è risultato più difficile del test 2005/2006 come si evince osservando la figura 4.51 che confronta gli *score* nei due test per gli stessi livelli di abilità; si nota infatti che la *Test Characteristic Curve* del test 2004/2005 è più a destra della curva del test del 2005/2006. Nonostante ciò il test del 2005/2006 ha ottenuto punteggi più bassi. Se i risultati delle rilevazioni fossero completamente affidabili e i questionari fossero stati somministrati nello stesso periodo dell'anno, si potrebbe desumere che tra il 2004/2005 e il 2005/2006 i livelli di apprendimento in IV elementare siano calati. In questa ipotesi, l'uso della metodologia in questione consentirebbe di tenere conto delle differenti difficoltà dei test e, l'eventuale aumento o diminuzione, potrebbe essere pienamente apprezzato attraverso la considerazione delle misure piuttosto che degli *score*. Però, considerando che il monitoraggio è avvenuto in due periodi differenti, e ipotizzando che i dati puliti del SNV siano completamente affidabili, le differenze tra le misure nelle due epoche (cioè tra i valori del 2005/2006 e i valori del 2004/2005) si possono interpretare come la crescita delle abilità in cinque mesi di apprendimento (tab. 4.7). In questo caso emerge la necessità della misurazione delle abilità, che quantifica su una scala a intervalli gli apprendimenti nei diversi momenti e permette il loro confronto (le operazioni di somma e sottrazione hanno un significato univoco), mentre gli *score* non forniscono alcuna informazione inerente alla variazione effettiva delle competenze, se non quella, puramente "ordinale", di indicare il passaggio da un valore medio più alto a un valore medio più basso (o viceversa) e, per uno scolaro o per una scuola, da una posizione all'interno della distribuzione a un'altra (poiché i test non sono omogenei tra loro il confronto tra gli *score* perde di significato, anche considerando i punteggi grezzi standardizzati).

La tabella 4.7 rappresenta gli indici sintetici (media, mediana, moda, *standard deviation* e quantili) degli *score* e delle misure delle due valutazioni del SNV calcolate su tre *data set*: 1) i database originali, 2) i database puliti dei casi di *cheating* e degli item con *misfit*, 3) i database puliti dei casi di *cheating* e degli item con *misfit* senza considerare i casi estremi (*zero score* e *perfect score*). La differenza tra lo *score* medio nel 2004/2005 e lo *score* medio nel 2005/2006 (calcolato sul *data set* originario) è pari a +1,7 (la differenza tra gli *score* mediani è pari a +2); la differenza tra le abilità medie nei due anni (calcolate sui *data set* puliti dai casi anomali) è pari a +0.82 (quella tra le abilità mediane a +0,88); si può allora dire che +0,82 indica grosso modo la variazione di conoscenze/competenze matematiche nell'arco di cinque mesi di scuola. La figura 4.45 rappresenta le distribuzioni delle abilità scolastiche medie nei due anni.

Un altro vantaggio derivante dall'applicazione del modello di Rasch è che esso fornisce una stima dell'errore di misura accanto alla stima dell'abilità e con la stima dell'errore di misura è possibile determinare gli intervalli di confidenza per le medie di misure e per le differenze tra medie di misure oltre che per le misure stesse. Per esempio, volendo determinare l'intervallo di confidenza per le differenze tra le medie di misure di una scuola i, calcolate in due momenti t_1 e t_2 distinti, si utilizza la statistica:

$$\frac{\overline{\theta}_{it_2} - \overline{\theta}_{it_1}}{\sqrt{\frac{\sum_{j=1}^{n_{jt_2}} SE_{jt_2}^2}{n_{jt_2}^2} + \frac{\sum_{j=1}^{n_{jt_1}} SE_{jt_1}^2}{n_{jt_1}^2}}} \tag{4.18}$$

dove $\overline{\theta}_{it_1}$ e $\overline{\theta}_{it_2}$ sono le medie della classe nei due momenti di rilevazione, n_{jt_1} e n_{jt_2} la numerosità della classe in t_1 e t_2 e SE_{jt_1} e SE_{jt_1} e SE_{jt_1} gli errori di misura del soggetto j appartenente alla classe i.

È interessante determinare quali sono state le variazioni a livello di scuola mettendo a confronto da un lato gli *score* e dall'altro le misure (usando il test 4.18) per valutare eventuali discrepanze. Nella tabella 4.8 si confrontano le differenze tra le medie degli *score* grezzi e le differenze tra le medie delle misure, sempre a livello scolastico. Nei database puliti sono state monitorate 5.980

scuole che hanno partecipato a entrambe le valutazioni; di queste 4.446 hanno registrato una differenza tra *score* medi negativa (per 4.302 anche la differenza tra misure medie è risultata negativa, per 139 la differenza tra misure medie non è stata significativamente diversa da zero e per 5 è stata significativamente maggiore di zero), 2 sole scuole hanno avuto lo score medio uguale nei due anni e 1.532 scuole hanno registrato una differenza tra *score* positiva (per 497 la differenza tra misure medie è risultata negativa, evidenziando che, in realtà, non c'è stato nessun miglioramento tra un'epoca e l'altra, per 638 la differenza non è risultata significativamente diversa da zero e per 397, invece, è stata significativamente maggiore di zero). Si evidenzia pertanto che le discrepanze risultano notevoli e l'uso delle misure è senza dubbio più opportuno poiché consente di tenere conto degli errori di misura.

4.4.2 La variabilità degli apprendimenti a livello regionale

Le ultime analisi riguardano i livelli di apprendimento a livello regionale. La figura 4.53 rappresenta le misure medie nel 2004/2005 e nel 2005/2006 in ciascuna regione, ordinate in base al livello del 2004/2005. Come era immaginabile aspettarsi avendo in mente le distribuzioni degli score (di cui le misure costituiscono una trasformazione logistica), sono le regioni meridionali (ad eccezione della Sardegna) ad aver registrato, in entrambi gli anni, i livelli più alti. Al Centro-Sud si sono registrate anche le maggiori variabilità delle misure come attesta la figura 4.54 che rappresenta le deviazioni standard dei livelli di conoscenza/competenza. A cosa sono imputabili queste variabilità? Se si scompone la varianza delle misure attraverso un modello multilivello con effetto scuola e classe del tipo:

$$misura_{ij\nu} = \mu + U_i + V_{ij} + e_{\nu ij} \tag{4.19}$$

con i la scuola, ij la classe j all'interno della scuola ij, νij lo scolaro ν della classe j nella scuola i, risulta che la varianza spiegata dalla scuola e dalla classe (fig. 4.55) è, in generale, molto alta con valori che, nel 2004/2005, superano

il 40% in tutte le regioni fatto salve per la Val d'Aosta (i valori più alti si registrano in Molise, Puglia, Sicilia , Campania, e Calabria, quelli più bassi in Val d'Aosta, Veneto, Lombardia, Piemonte e Friuli-Venezia Giulia). Di questa varianza spiegata dalla classe e dalla scuola, la quota spiegata dalla scuola (fig. 4.56) è inferiore a quella spiegata dalla classe (si va dal 40% delle Marche, Campania, Sicilia e Puglia al 20% o meno della Val d'Aosta, Liguria, Trentino e Friuli-Venezia Giulia). E ciò fa pensare che, nel bene (efficacia) o nel male (*cheating*), ciò che succede a livello di classe abbia un ruolo molto importante, maggiore che a livello di scuola.

È interessante, infine, constatare come i valori medi delle misure diminuiscano all'aumentare della variabilità all'interno della classe (fig. 4.57), che sembrerebbe suggerire che nelle classi dove c'è maggior eterogeneità si impara meno. Infatti se si prendono in considerazione le differenze delle misure tra i due anni (una sorta di valore aggiunto del processo educativo, tenendo presente, però che si tratta di coorti di scolari diverse) anziché i livelli di abilità, l'ordine delle regioni è, parzialmente, invertito, con alcune regioni del Nord (Emilia Romagna, Veneto e Val d'Aosta assieme a Marche, Tosca e Umbria) che evidenziano le performance migliori e alcune regioni meridionali (Basilicata, Sicilia e Sardegna, assieme a Lazio e Piemonte) che ottengono i risultati meno brillanti.

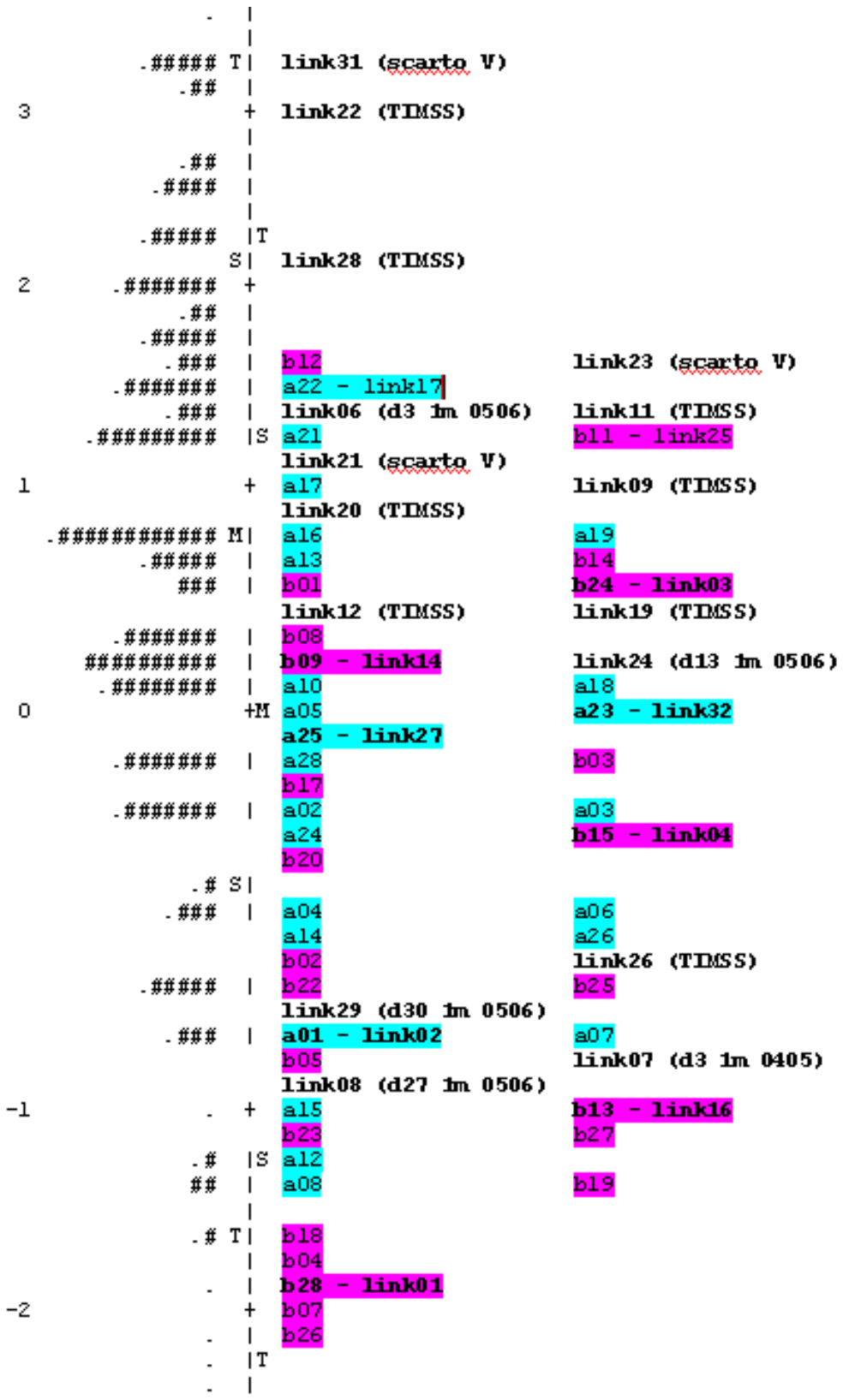


Figura 4.47: La scala di misurazione con i 64 item

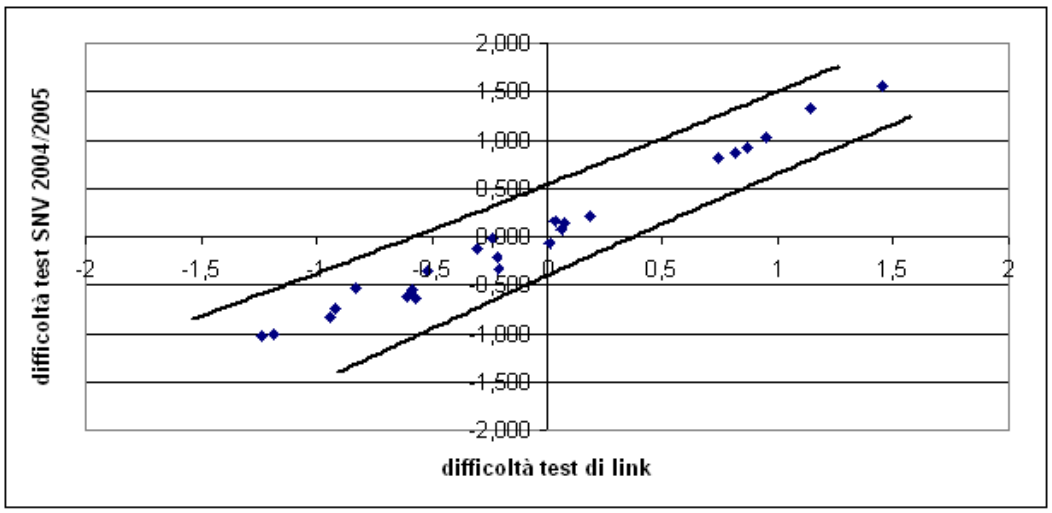


Figura 4.48: Difficoltà degli item del SNV 2004/2005 stimate prima e dopo la costruzione della scala

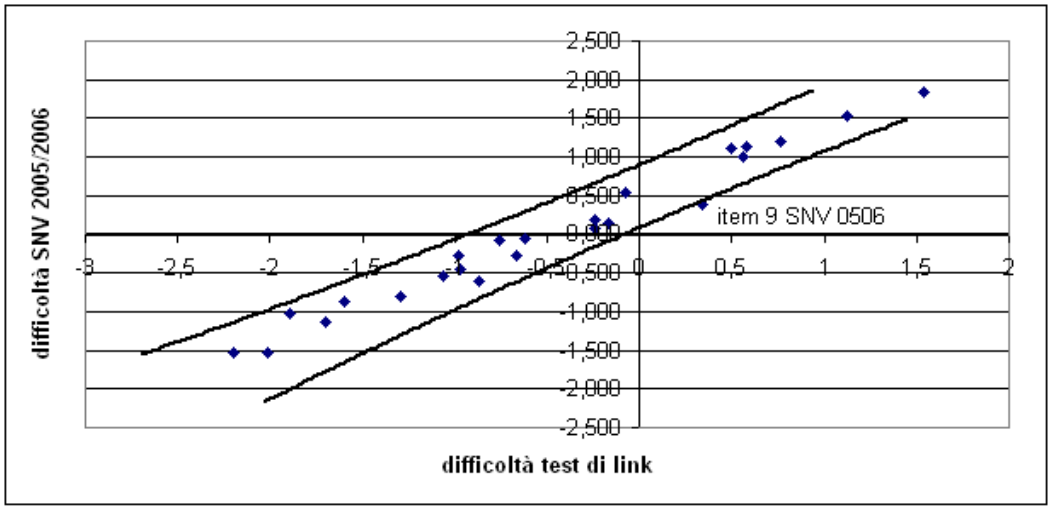


Figura 4.49: Difficoltà degli item del SNV 2005/2006 stimate prima e dopo la costruzione della scala

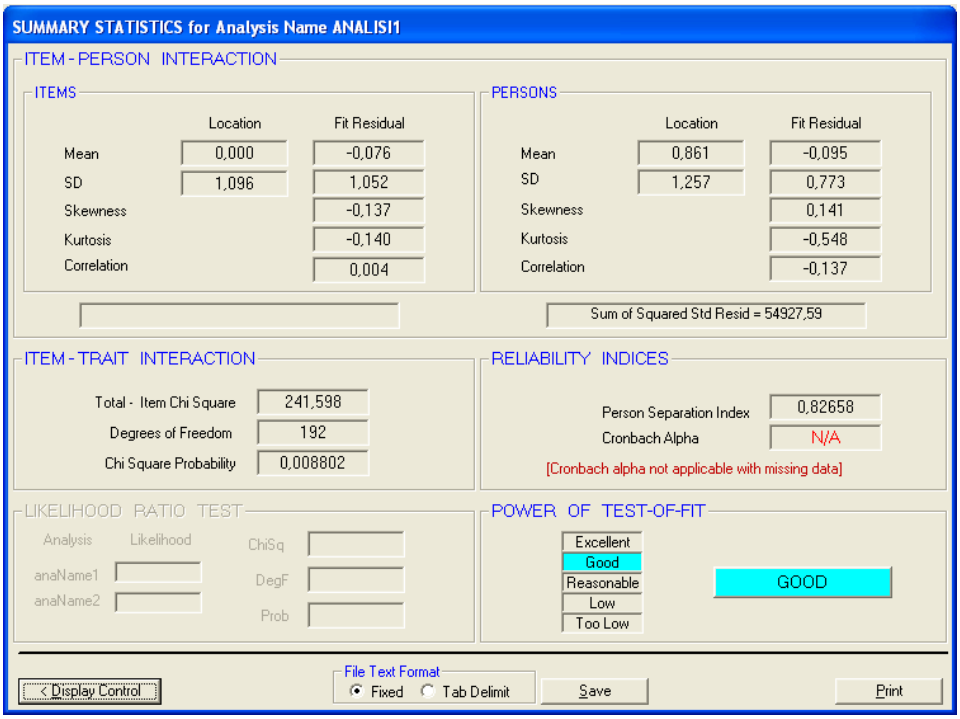


Figura 4.50: Summary Statistics della scala

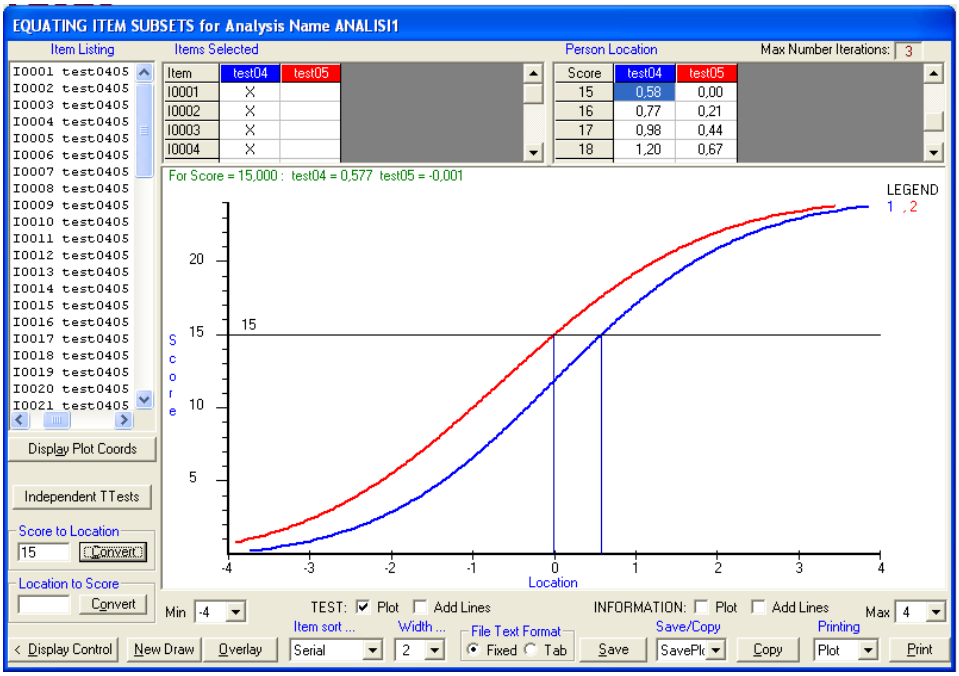


Figura 4.51: I due test riportati sulla stessa scala

	Dati grezzi		Dati puliti No cheat./It misf.		Dati puliti senza extreme score	
	2004/2005	2005/2006	2004/2005	2005/2006	2004/2005	2005/2006
N	480.897	491.612	415.638	423.961	394.377	412.611
Media	20,1	18,4	17,0	15,2	16,7	15,0
Mediana	21	19	18	15	17	15
Moda	27	27	21	16	21	16
sd	5,6	6,1	4,6	4,9	4,4	4,7
I quartile	16	14	14	12	14	12
II quartile	21	19	18	15	17	15
III quartile	25	24	21	19	20	19
Media	-	-	1,2	0,4	1,0	0,3
Mediana	-	-	1,1	0,1	0,9	0,1
Moda	-	-	2,1	0,3	2,1	0,3
sd	-	-	1,4	1,4	1,1	1,2
I quartile	-	-	0,3	-0,5	0,3	-0,5
II quartile	-	-	1,1	0,1	0,9	0,1
III quartile	-	-	2,1	1,1	1,7	1,1

Tabella 4.7: Score e misure

		measure diff.		
		signif. < 0	non signif. $\neq 0$	signif. > 0
	< 0	4302	139	5
score diff.	0	1	1	0
	> 0	497	638	397

Tabella 4.8: Differenze tra score medi e differenze tra misure medie a livello di scuola

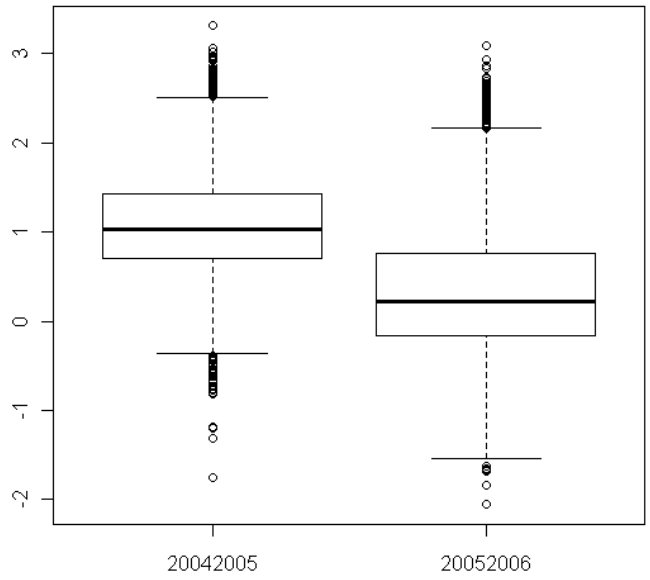


Figura 4.52: Boxplot delle abilità medie scolastiche

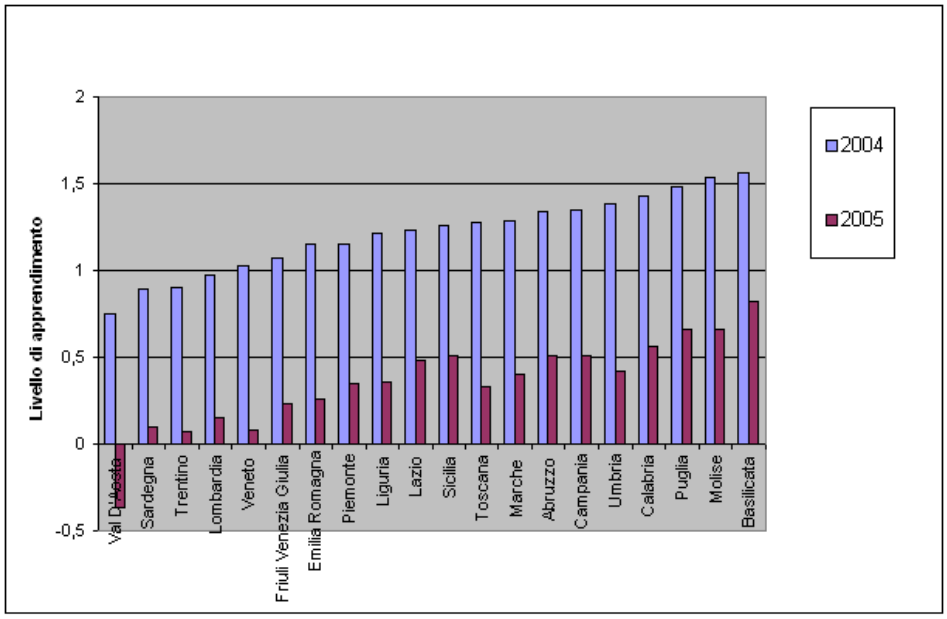


Figura 4.53: Livelli medi di apprendimento regionali

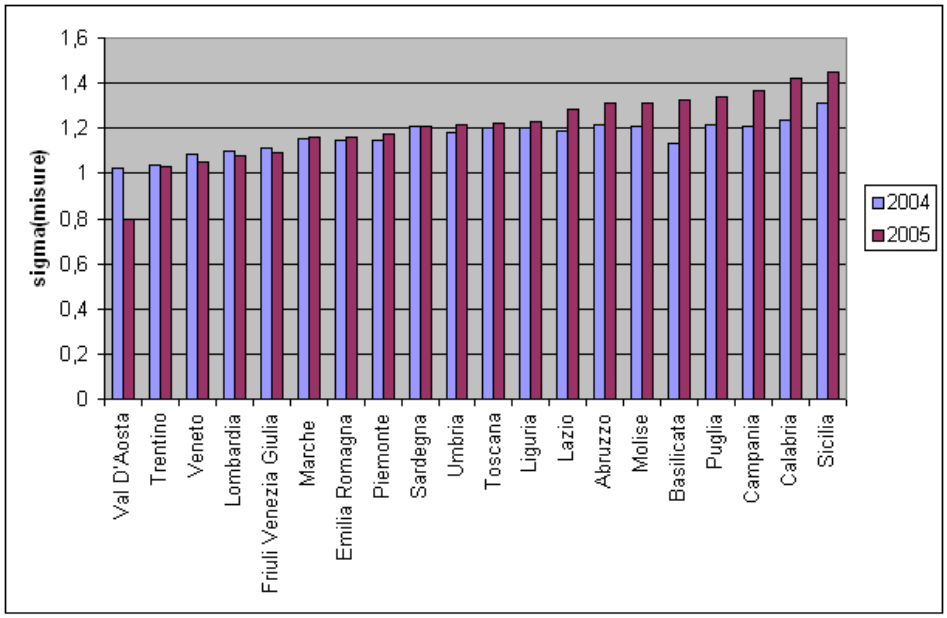


Figura 4.54: Deviazione standard degli apprendimenti

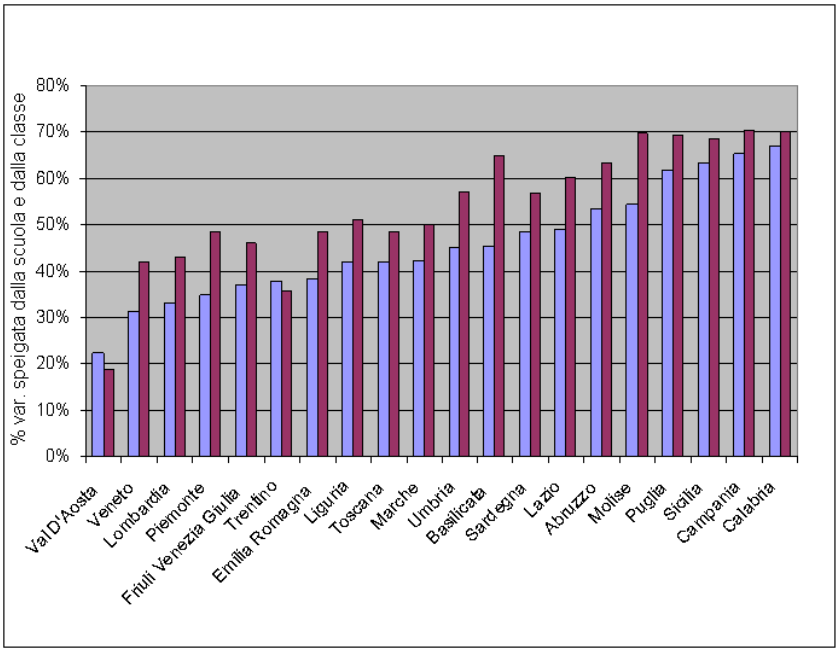


Figura 4.55: Varianza spiegata dalla scuola e dalla classe

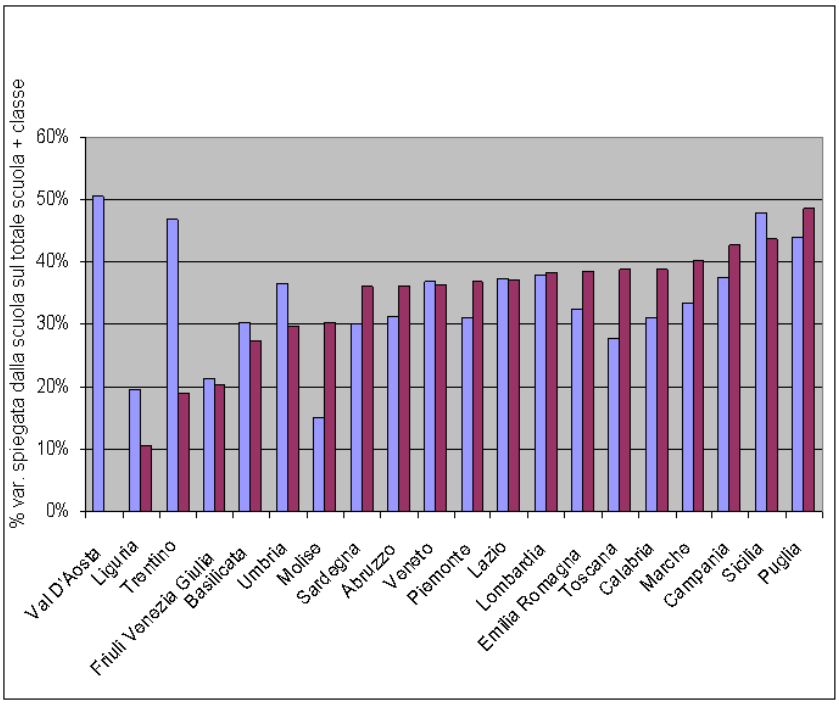


Figura 4.56: Varianza spiegata dalla scuola sul totale della varianza spiegata dalla scuola e dalla classe

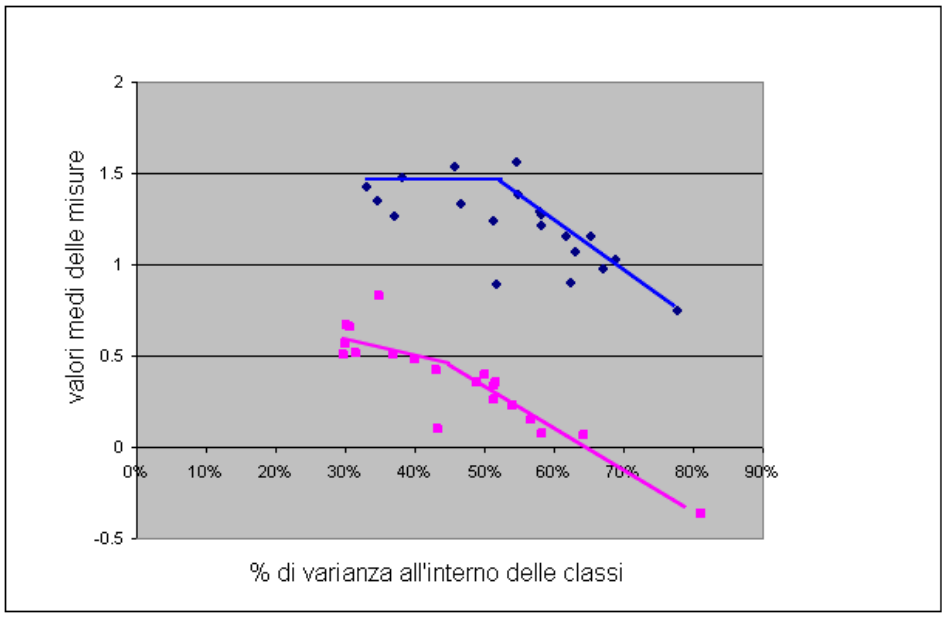


Figura 4.57: Valori medi di apprendimento rispetto alla percentuale di varianza all'interno delle classi

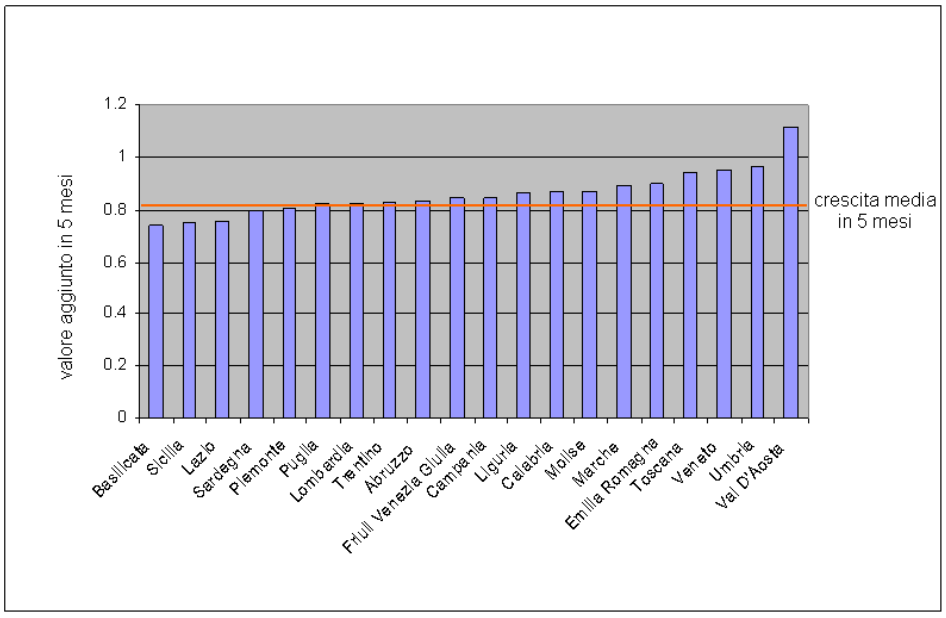


Figura 4.58: Valori aggiunto regionale in cinque mesi di apprendimento

Bibliografia

1. Andersen E. B. (1973a). A goodness of fit test for the Rasch Model. Psychometrika, Vol. 38, pp. 123-140.

2. Andersen E. B. (1973b). Conditional inference for multiple choice questionnaires. British Journal of Matematical and Statistical Psychology, Vol. 26, pp. 31-44.

3. Andersen E. B. (1973c). Conditional inference and models for measuring. Copenhagen: Mentalhygiejnisk Forlag.

4. Andersen E. B. (1977). Sufficient Statistics and latent trait model. Psychometrika, Vol. 42, pp. 69-81.

5. Andersen E. B., Olsen L. W. (2001). The life of Georg Rasch as a mathematician and as a statistician. In Boomsma A., van Duijn M. A. J., Snijders T. A. B. (Eds.). Essays on item response theory, *Lecture Notes in Statistics*, Vol. 157, pp. 3-24. New York: Springer Verlag.

6. Andrich D. (1978a). A rating formulation for ordered response categories. Psychometrika, Vol. 43, pp. 357-374.

7. Andrich D. (1978b). Relationships between the Thurstone and Rasch approaches to item scaling. Applied Psychological Measurement, Vol. 3, pp. 446-460.

8. Andrich D. (1978c). Scaling attitude items constructed and scored in the Likert tradition. Educational and Psychological Measurement., Vol. 38, pp. 665-680.

9. Andrich D. (1982a). An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman Scale response pattern. Educational Research and Perspectives, UWA, Vol. 9, n. 1, pp. 95-104.

10. Andrich D. (1982b). An extension of the Rasch model for ratings providing both location and dispersion parameters. Psychometrika, Vol. 47, pp. 105-113.

11. Andrich D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. Brandon-Tuna N., Sociological Methodology, San Francisco, Jossey-Bass, Ch. 2, pp. 33-80.

12. Andrich D. (1988a). Rasch Models for Measurement. Series: Quantitative Applications in the Social Sciences. CA: Sage Press Newbury Park.

13. Andrich D. (1988b). A general form of Rasch's Extended Logistic Model for Partial Credit Scoring. Applied Measurement in Education, Vol. 1, N. 4, pp. 363-378.

14. Andrich, D. (2009). The change of metric in applying the Rasch model to quantify and account for response dependence between items. To be submitted to Psychometrika.

15. Andrich D. and Kline P. (1981). Within and among population item fit with the simple logistic model. Educational and Psychological Measurement, Vol. 41, pp. 35-48.

16. Andrich D., Luo G. (2003). Conditional pair-wise estimation in the Rasch Model for ordered response categories using principal components. Journal of Applied Measurement, Vol. 4, n. 3, pp. 205-221.

17. Andrich D., Hagquist C. (2004). Detection of Differential Item Functioning using Analysis of Variance. Paper presented at the Second International Conference on Measurement in Helath, Education, Psychology and Marketing: Developments with Rasch Models. Murdoch University, Perth.

18. Andrich D., Sheridan B., LUO G. (2000). RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models. Perth: Murdoch University.

19. Barbanelli C., Natali E. (2005). I Test Psicologici: Teorie e Modelli Psicometrici. Carocci editore.

20. Binet A. e altri (1894). Introduction à la psychologie expérimentale (Introduzione alla psicologia sperimentale).

21. Bond T. G., Fox C. M. (2007). Applying the Rasch Model, Fundamental Measurement in Human Sciences (2nd ed.). Lawrence Erlbaum Associates, Mahwah, New Jersey.

22. Choppin B. (1983). A fully conditional estimation procedure for Rasch Model parameters. Report n. 196. Los Angeles: University of California, Graduate School of Education Center for the Study of Evaluation.

23. Cristante F., Mannarini S. (2004). Misurare in psicologia: Il Modello di Rasch. Laterza.

24. Crocker L. M., Algina J. (1986). Introduction to classical and modern test theory. CA: Wadsworth, Belmont.

25. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, Vol. 16, pp. 297-334.

26. De Leeuw J., Verhelst N. D. (1986). Maximum likelihood estimation in generalized Rasch Models. Journal of Educational Statistics, Vol. 11, pp. 183-196.

27. Fischer G. H., Scheiblechner H. (1970). Algorithmen und programme für das probabilistischen testmodell von Rasch (Algorithms and programs for the probabilistic test model of Rasch). Psychologische Beiträge, Vol. 12, pp. 23-32.

28. Fischer G. H. (1974). Einfüehrung in die Theorie psychologischer Tests (Introduction to mental test theory). Bern: Huber.

29. Fischer G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. Psychometrika, Vol. 46, pp. 59-77.

30. Fischer G. H., Molenaar I. W. (1995). Rasch Models. Foundations, Recent Developments, and Applications. Springer Verlag.

31. Glas C. A. W. (1988a). The derivation of some tests for the Rasch model from the multinomial distribution. Psychometrika, Vol. 53, pp.525-546.

32. Glas C. A. W. (1988b). The Rasch model and multi-stage testing. Journal of Educational Statistics, Vol. 13, pp. 45-52.

33. Glas C. A. W. (1989). Contributions to estimating and testing Rasch models. (Doctoral Thesis). Enschende: University of Twente.

34. Glas C. A. W., Verhelst N. D. (1989). Extensions of the partial credit model. Psychometrika, Vol. 54, pp. 635-659.

35. Gori E., Sanarico M., Plazzi G. (2005). La valutazione e la misurazione nelle scienze sociali: oggettività specifica, statistiche sufficienti e modello di Rasch. Non Profit, Vol. 3, pp. 605-643.

36. Gori E., Vittadini G. (1999). La valutazione dell'efficienza ed efficacia: definizioni, problemi e metodi, in "Qualità e valutazione nei servizi di pubblica utilità", (Gori-VITTADINI eds.) ETAS, serie Gestione d'Impresa-Direzione.

37. Gruijter de N. M., Van der Kamp J. Th. (2003). Statistical Test Theory for Education and Psychology. D. N. M. de Gruijter & L. J. Th. Van der Kamp.

38. Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. The Annals of Statistics, Vol. 5, pp. 815-841.

39. Hambleton R. K., Swaminathan H. (1985). Item response theory: Principles and applications. Kluwer-Nijhoff, Boston.

40. Hambleton R. K., Swaminathan H., Rogers H. J. (1991). Fundamentals of Item Response Theory. CA: Sage Press Newbury Park.

41. Hosmer David W., Lemeshow S. (2000). Applied Logistic Regression (2nd ed.). John Wiley and Sons.

42. Humphry S. M. (2005). Maintaining a common arbitrary unit in social measurement. Ph.D. Thesis: http://www.lib.murdoch.edu.au/adt/browse/view/adt-MU2005830.95143

43. Humphry S. M. (2006). The impact of differential discrimination on vertical equating. (ARC report)

44. Humphry S. M., Andrich D. (2007). Understanding the unit in the Rasch Model. The University of Western Australia.

45. Karabatsos G. (2000). A Critique of Rasch residual fit statistics. Journal of Applied Measurement, 1(2), pp. 152-176.

46. Kelley, T. L. (1947). Fundamentals of statistics, Cambridge: Harvard University Press.

47. Kuder, G. F., Richardson M. W. (1937). The theory of the estimation of test reliability. Psychometrika, Vol. 2, pp. 151-60.

48. Levitt S. D. (2002). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. Berkeley Olin Program in Law and Economics, Working Paper Series 1078, Berkeley Olin Program in Law and Economics.

49. Linacre J. M. (1989). Many-facet Rasch measurement. Chicago: MESA Press.

50. Linacre J. M. (1998). Detecting multidimensionality: Which residual data-type works best. Journal of Outcome Measurement, Vol. 2, pp. 266-283.

51. Lynch, B. K., McNamara, T. F. (1998). Using g-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. Language Testing, Vol. 15, n. 2, pp. 158-80.

52. Lord F. M., Novick M. R. (1968). Statistical theory of mental test scores. Addison Wesley, Reading, Mass.

53. Lord F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter model. Educational and Psychological Measurement, Vol. 28, pp. 989-1020.

54. Lord F.M. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. (Research Report RB-75-33). Princeton: ETS.

55. Luo G., Seow A., Chin C. L. (2001). Linking and anchoring techniques in test equating using the Rasch Model. Nanyang Technological University, Singapore.

56. Marais I., Andrich D. (2008). Formalising dimension and response violations of local independence in the unidimensional Rasch model. Journal of Applied Measurement, Vol. 9, n. 3, pp. 200-215.

57. Marchi Mario (2004). Quadro di riferimento per le Prove di Valutazione in Matematica. INVALSI.

58. Masters G. N.(1982). A Rasch Model for Partial Credit Scoring. Psychometrika, Vol. 47, n. 2, pp. 149-174.

59. Michell, J. (1997). Quantitative science and the definition of measurement in psychology. British Journal of Psychology, Vol. 88, n. 3, pp. 355-383.

60. Molenaar I. W. (1983). Some improved diagnostics for failure in the Rasch model. Psychometrika, Vol. 48, pp. 49-72.

61. Perelli D'Argenzio M. P., (2006). La valutazione esterna degli apprendimenti: le prove di valutazione INVALSI. Le prove di matematica. L'insegnamento della matematica e delle scienze integrate, Vol. 29, n. 1, pp. 31-46.

62. Rasch G. (1960). Probabilistic models for some intelligence and attainment tests (expanded edition). Copenhagen: The Danish Institute of Educational Research.

63. Rasch G. (1961). On general laws and the meaning of measurement in psychology. Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Theory of Probability, Vol. IV, pp. 321-333. Berkeley: University of California Press.

64. Rasch G. (1977). On Specific Objectivity: An attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy, Vol. 14, pp. 58-94.

65. Smith E. V. Jr (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement, 3(2), 205-231.

66. Smith R. M., Schumacker R. E., Bush M. J. (1998). Using item mean squares to evaluate fit to the Rasch Model. Journal of Outcome Measurement, 2(1), pp. 66-78.

67. Smith R. M. (1988). The distributional properties of Rasch standardized residuals. Educational and Psychological Measurement, Vol. 48, pp. 657-667.

68. Smith R. M. (1991). The distributional properties of Rasch item fit statistics. Educational and Psychological Measurement, Vol. 51, pp. 541-565.

69. Smith R. M. (1992). Assessing unidimensionality for the Rasch rating scale model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco (CA).

70. Smith R. M. (1994). A comparison of the power of Rasch total and between-item fit statistics to detect measurement disturbances. Educational and Psychological Measurement, Vol. 54, n. 1, pp. 42-55.

71. Smith R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. Structural Equation Modeling, Vol. 3, pp. 25-40.

72. Smith R. M. (2000). Fit analysis in latent trait measurement models. Journal of Applied Measurement, 1(2), pp. 199-218.

73. Smith R. M., Miao C. Y. (1994). Assessing unidimensionality for Rasch measurement. Objective Measurement: Theory into Practice, Vol. 2, pp. 316-327.

74. Spearman C. (1904a). The proof and measurement of the association between two things. American Journal of Psychology, Vol. 15, pp. 72-101.

75. Spearman C. (1904b). General intelligence objectively determined and measured. American Journal of Psychology, Vol. 15, pp. 201-293.

76. Stocking M.L. (1989). Empirical estimation errors in item response theory as a function of test properties. Research Report RR-89-5, Princeton: ETS.

77. Swaminathan H. (1983). Parameter estimation in item response models, in Hambleton R. (Ed.), Applications of Item Response Theory, "Vancouver", BC: Educational Research Institute Of British Columbia, pp. 24-44.

78. Taylor J. R. (1982). An introduction to error analisys. The study of uncertainties in physical measurements. University Science Books, Mill Valley (CA).

79. Tesio, L. (1995). Independence: the core and currency of functional assessment in rehabilitation medicine. Acta Gerontol, Vol. 45, pp. 133-136.

80. Thurstone L. L. (1931). The reliability and validity of tests. Edwards Brothers, Ann Arbor.

81. Van der Linden W. J., Eggen T. J. H. M. (1986). An empirical Bayesian approach to item banking. Applied Psychological Measurement, Vol. 10, pp. 345-354.

82. Van den Wollenberg A. L. (1982). Two new test statistics for the Rasch model. Psychometrika, Vol. 47, pp. 123-139.

83. Verhelst N. D., Eggen T. J. H. M. (1989). Psychometrische en statistische aspecten van peilingsonderzoek (Psychometric and statistical aspects of assessment research). (PPON-rapport, 4) Arnhem: CITO.

84. Vidoni D., Falzetti P., Plazzi G. (2009). Le rilevazioni nazionali condotte dall'INVALSI. Questioni da risolvere e necessità di un protocollo per la correzione dei dati. Presentato al Convegno della Banca d'Italia tenutosi il 6 marzo 2009 a Roma.

85. Wingersky M.S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models, in Hambleton R.K., ed. Applications of Item Response Theory.

86. Wright B.D. (1968). Sample-free test calibration and person measurement. In "Proceedings of the 1967 Invitational Conference on Testing Problems". Princeton, N.J.: Educational Testing Services.

87. Wright B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, Vol. 14, pp. 97-116.

88. Wright B. D., Stone M. H. (1979). Best Test Design. MESA.

89. Wright B. D., Masters G. N. (1982). Rating Scale Analysis. MESA.

90. Wright B. D., Panchapakesan N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, Vol. 29, pp. 23-48.

91. Zimmerman D. W. (1976). Test theory with minimal assumptions. Educational and Psychological Measurement, Vol. 36, pp. 85-96.

92. Zimmerman D. W., Williams R. H. (1977). The theory of test validity and correlated errors of measurement. Journal of Math. Psychology, Vol. 16, pp. 135-152.

93. Zwinderman A. H. (1995). Pairwise parameter estimation in Rasch Models. Applied Psychological Measurement, Vol. 19, n. 4, pp. 369-375.

94. Council Conclusions on a strategic framework for European cooperation in education and training ("ET 2020"). 2941th Education, Youth and Culture Council meeting. Brussels, 12 May 2009