DOTTORATO DI RICERCA IN
SCIENZE STORICO-SOCIALI
METODOLOGIA DELLE SCIENZE SOCIALI

CICLO XXVII

COORDINATORE PROF. MARCO BONTEMPI

La misurazione nella ricerca sociale: il modello di Rasch e la teoria della misurazione additiva congiunta

Settore Scientifico Disciplinare SPS/07

Dottorando
Dott. LA SALA DIEGO

_____
(firma)

Tutore
Prof. LANDUCCI SANDRO

_____
(firma)

Coordinatore
Prof. BONTEMPI MARCO

_____
(firma)

Anni 2012/2014

A Brvno

Indice generale

INTRODUZIONE

L'oggetto principale della presente tesi è lo studio del modello di Rasch (sullo sfondo più generale dei modelli probabilistici della *Item Response Theory*) come strumento per la costruzione di scale a intervalli capaci di rappresentare dimensioni attitudinali latenti d'individui. Il tentativo proposto è quello di svolgere la riflessione sulle sue caratteristiche, i suoi assunti e le sue proprietà matematiche, in continua relazione con i più generali concetti di misurazione e di quantità continue misurabili. Il taglio scelto è quello di una disamina che tenga insieme, nel modo più organico possibile, la sfera tecnica e formale del modello con quella epistemologica relativa alle condizioni di conoscibilità di atteggiamenti e opinioni individuali.

L'obiettivo è di avanzare un'interpretazione e un giudizio sull'ipotesi che il modello di Rasch, seppure non come unico esempio possibile, rappresenti uno strumento capace di raggiungere, nell'ambito delle scienze umane, lo scopo della misurazione, in un modo analogo a quello di scienze più avanzate, o se quantomeno permetta di imboccare una strada simile, come sostenuto da alcuni studiosi. In questo senso, oltre all'approfondimento delle proprietà tecniche del modello e del loro rapporto con le più ampie questioni epistemologiche sollevate dal particolare oggetto di ricerca (gli atteggiamenti), sarà indagato anche il rapporto tra il modello e gli assiomi della Teoria della misurazione additiva congiunta (ACM), sviluppata per individuare e testare nel modo più generale possibile le condizioni necessarie ad accertare la natura additiva e continua delle strutture empiriche (anche di quelle non estensive del mondo fisico) e quindi le condizioni di possibilità di una misurazione su scale a intervalli. Per ricostruire questo quadro d'insieme e renderlo organico, sarà necessario affrontare una parte significativa del dibattito, anche molto recente, su questo tema, dando un contributo per riorganizzarlo attorno ad alcuni assi principali, selezionandone gli aspetti fondamentali.

Nella Parte Prima del testo si affronterà il concetto di *misurazione* in generale e nelle scienze umane, non per darne una definizione univoca o difenderne una in particolare, né con l'ambizione di ricostruire in modo completo i termini di un dibattito secolare sul suo significato e sulle sue fattispecie. Piuttosto, l'intenzione è di individuare e dare conto di alcuni filoni tematici relativi alla questione, in termini di posizioni filosofiche e sviluppo storico, al fine di fornire una prima, generale illustrazione della sua complessità (capitolo 1).

Sarà quindi introdotto brevemente il concetto di *scaling* nelle scienze umane, come insieme di procedure per costituire scale di misurazione e costruire misure, seguito da una sintetica selezione di criteri di classificazione e identificazione di vari modelli (capitolo 2). Questo rappresenterà la premessa per l'illustrazione di tre approcci classici allo *scaling*: quello di Thurstone, quello di Coombs e quello di Guttman (capitolo 3), tutti in qualche modo connessi con la discussione successiva.

Nella Parte Seconda introdurremo la Teoria del Tratto Latente e i modelli probabilistici della *Item Response Theory* (capitolo 4), per poi entrare nel dettaglio dei modelli unidimensionali per risposte dicotomiche (capitolo 5): i loro assunti comuni, il significato dei parametri delle loro funzioni caratteristiche, la natura delle curve del test e i caratteri della capacità informativa degli *item* e dei test.

Mostreremo i modelli di base a tre e due parametri, fornendone un'interpretazione (capitolo 6). Lasceremo poi spazio al modello di Rasch, illustrandone la derivazione, discutendone le importanti proprietà e caratteristiche strutturali e analizzando criticamente le sue potenzialità per i processi di misurazione (capitolo 7).

La Parte Terza è dedicata alla discussione del fondamento del modello di Rasch come fattispecie di misurazione in grado di produrre scale a intervalli e sulla dimostrazione della natura quantitativa delle strutture empiriche degli attributi che intenderebbe misurare. Nel capitolo 8 sarà ripresa questa tematica e introdotta l'ipotesi, nota in letteratura, che il modello di Rasch rappresenti una fattispecie probabilistica di misurazione additiva congiunta e sia quindi in grado di testare la natura degli attributi. Saranno presentati e commentati i presupposti teorici e gli assiomi fondamentali dell'ACM (capitolo 9), per poi confrontarli in modo sistematico con le caratteristiche del modello di Rasch e mostrare che la probabilità di superare un *item*, data la sua difficoltà e la posizione di un soggetto sulla dimensione latente, rappresenti una forma matematica di struttura additiva congiunta (capitolo 10). Nel successivo capitolo 11, saranno invece presi in considerazione alcuni elementi di differenza o debolezza nel legame tra modello di Rasch e ACM, che spingono a guardare in modo problematico il loro rapporto.

Discuteremo poi due posizioni che tendono a negare in modo più netto il legame tra ACM e modello di Rasch, esemplificate dagli argomenti di altrettanti autori. La pri-

ma è quella espressa da Kyngdon, da un'ottica legata a una supposta interpretazione rigorosa dei principi della Teoria rappresentazionale della misurazione (capitolo 12). Questo punto di vista sarà approfondito criticamente, alla luce di una concezione più articolata della natura dei concetti empirici.

La seconda posizione è quella di Michell (capitolo 13), ispirata a un'ottica realista empirica e a una concezione tradizionale della misurazione, incentrata sulla polemica contro un'ipotetica tendenza al non accertamento della natura quantitativa e continua degli attributi mentali, data per scontata da molti ricercatori. Il quattordicesimo e ultimo capitolo è dedicato all'illustrazione di quello che si definisce come "paradosso dei modelli di Guttman e Rasch", che metterebbe in discussione la possibilità logica che il modello di Rasch possa produrre scale a intervalli. Analizzeremo due tipi di critiche al paradosso, provando a capire se esse siano o no in grado di smentirlo e neutralizzarne le conclusioni. Infine presenteremo una diversa giustificazione del rapporto tra probabilità e precisione, imperniata sull'analogia tra l'utilizzo della probabilità nella teoria della meccanica quantistica e nei modelli psicometrici, discutendo le differenze profonde nei rispettivi quadri epistemologici.

Nelle conclusioni faremo il punto della riflessione generale compiuta attraverso il testo, individuando i risultati principali, i punti specifici lasciati aperti e i possibili sviluppi futuri del campo cognitivo indagato, sulla base di quanto raccolto e discusso nelle tre parti della tesi.

PARTE PRIMA

Misurazione e *scaling*: qualche concetto fondamentale e alcuni esempi

1 Il concetto di misurazione: brevi cenni generali

Il tema della *misurazione*, dei suoi procedimenti e dei suoi requisiti è da tempo immemore, in tutta la sua complessità, dibattuto in ambito filosofico e scientifico, soprattutto nel campo della ricerca sociale e psicologica, nel quale è stato importato, più di un secolo fa, da altri settori disciplinari. Infatti, i procedimenti di misurazione comunemente intesi sono nati e si sono sviluppati prima di tutto e principalmente nella pratica delle scienze fisiche, dove la quantificazione di grandezze empiriche ha svolto, e svolge, un ruolo fondamentale nel grande progresso e nei successi raggiunti in quell'area della conoscenza. Le grandi conquiste cognitive ivi ottenute ne hanno fatto ben presto il modello, il paradigma per tutte le scienze: sin dall'Ottocento, le tradizioni di pensiero che, in misura maggiore o minore, si sono ispirate ai principi del filone che possiamo indicare con l'etichetta generica di "positivismo", hanno sostenuto che ogni impresa scientifica degna di questo nome avrebbe dovuto ripercorrere, nei metodi e nell'epistemologia, il cammino delle scienze fisiche. Anche le scienze umane (psicologia, sociologia ecc.) non avrebbero potuto esimersi dal riprodurre nei loro rispettivi campi gli schemi adottati così proficuamente in fisica, in chimica e nelle altre scienze naturali. In quest'ottica, si comprende quale rilevanza strategica assuma la capacità di *quantificare* e, se possibile, di *misurare* i costrutti in esame, di trasformare in *numeri* gli *stati empirici* della realtà e di utilizzare i potenti strumenti della *matematica* per descrivere le *connessioni* tra fenomeni della realtà[1].

Nel senso comune, nella vita di tutti i giorni, la misurazione è identificata con l'idea di una *stima numerica* di grandezze quantitative. Un autorevole dizionario della lingua italiana, Treccani, definisce "misurazione" come "l'operazione del misurare, consistente nel confrontare una determinata grandezza fisica con la sua unità di misura, allo scopo di determinare il valore (o misura) della grandezza stessa". Si può facilmente concordare, sulla base dell'esperienza quotidiana, che nella comunicazione pubblica l'impresa scientifica stessa, *tout court*, è talvolta assimilata quasi completamente con il

[1] Tra gli altri, fu il grande Galileo Galilei, secoli prima, a incoraggiare questo connubio profondo tra matematica e scienze naturali, arrivando a sostenere che la prima fosse il "linguaggio" stesso con cui è stato scritto il grande "libro della Natura" [Frova e Marenzana 1998]. Su questa scia ritroviamo praticamente tutti i grandi scienziati di epoca storica moderna, a partire da Newton.

compito di misurare *qualcosa*, di stabilire grandezze da inserire in uno schema più generale di rapporti e di leggi di natura matematica.

Che le operazioni di rilevazione[2] delle dimensioni attitudinali, valoriali [Marradi 2005] e psicologiche degli individui possano, sotto certe condizioni e in alcune circostanze, essere assimilate a vere e proprie forme di misurazione, è argomento da sempre molto controverso. I motivi del dibattere sono molti e di varia natura (sia teorici che pratici). Pur con una certa doverosa cautela, ci pare che un modo di sintetizzare la questione, in ultima istanza, sia possibile attraverso un grande nodo centrale: la *questione ontologica*, cioè la diversità dell'oggetto di ricerca delle scienze naturali e di quelle sociali, da cui discendono di riflesso importanti questioni epistemologiche e metodologiche [es. Hughes e Sharrock 1997; Sparti 2002]. Non è forse un caso che questi temi si ripresentino ciclicamente nel dibattito delle scienze sociali, sin dai loro albori, senza riuscire mai a esaurirsi completamente: i costrutti mentali che vorremmo valutare[3] negli individui umani, non hanno il carattere osservabile e manipolabile di quelli fisici o almeno il loro livello di strutturazione empirica, ma sono proprietà latenti, non osservabili, ipotizzabili ma testabili solo indirettamente. Da qui le controversie epistemologiche e le difficoltà metodologiche che da sempre affrontano i ricercatori sociali intenzionati a occuparsi di atteggiamenti, opinioni, valori.

Nei molti decenni di storia della riflessione sulla ricerca psicologica e sociologica, è riscontrabile uno spettro molto ampio di posizioni sul tema: si va dal ritenere la misurazione dei costrutti mentali logicamente *impossibile* anche solo in via di principio, al sostenere invece che quelle sociali e dell'uomo possano definirsi "scienze" allo stesso modo di quelle "esatte" e affrancarsi dalla mera speculazione filosofica *solo* nel grado in cui siano capaci di produrre misurazioni empiriche e leggi. Non è necessario ai nostri scopi ricostruire dettagliatamente le ragioni di ogni punto di vista nel dibattito epistemologico[4]. E' tuttavia utile, ai fini della presente ricerca, toccare almeno in parte la complessità teorica dell'argomento e riflettere in via preliminare su alcune definizioni del

[2] Il termine "rilevazione" per significare in generale anche le operazioni di stima delle caratteristiche latenti dei soggetti è adottato, per esempio, da Marradi [1980] e Pitrone [2009], come alternativa più generale al termine "misurazione".

[3]Cannavò [1999] propone il termine "valutazione" per nominare le operazioni di cui alla nota precedente.

[4] I punti fondamentali per la nostra trattazione saranno successivamente ripresi e adeguatamente approfonditi nella Parte Terza.

concetto di misurazione, tenendo conto dei loro presupposti teorici e dei nodi attorno ai quali via via s'incardina il dibattito. Ciò servirà a rilevare, intanto, un elemento che persisterà lungo tutta la riflessione successiva sui modelli di *scaling* in generale e su quello di Rasch in particolare: il peso dirimente delle scelte definitorie e degli assunti adottati per giustificare le operazioni di rilevazione, aspetti molto difficilmente aggirabili in tutte le fasi della ricerca scientifica in campo sociale.

1.1 Il paradigma classico della misurazione

Michell [2007], descrivendo la concezione "classica" della misurazione nelle scienze fisiche, scrive[5] che "una *misura* di una grandezza di un attributo quantitativo è una stima del rapporto tra tale grandezza e qualsiasi altra dello stesso attributo sia presa come unità di misura"[6]. Proviamo a isolare gli aspetti rilevanti della definizione. Innanzitutto, la misurazione non riguarda gli oggetti in sé, bensì i loro *attributi* (o proprietà, o caratteri, o tratti) e le relazioni tra loro. Secondo, la quantità specifica (la *grandezza*) dell'attributo di un oggetto è accertabile tramite il confronto empirico tra più oggetti che possiedono tale attributo. Terzo, una certa grandezza è fissata come *unità* di misura, termine di confronto e di paragone. Infine, la misura è un *numero* puro, che rappresenta il rapporto tra la grandezza da misurare e l'unità di misura, secondo una logica additiva. L'esempio paradigmatico di quest'accezione di misurazione per grandezze concatenabili è la *lunghezza* di un segmento.

Si utilizza in questa definizione una distinzione adoperata già da Aristotele riguardo alle *quantità*: quando esse sono enumerabili, discrete, egli parla di *pluralità*, il cui accertamento avviene attraverso l'operazione del conteggio[7]; parla invece di *grandezze* misurabili quando la quantità è definita come continua, caratterizzata cioè da un

[5] Da questo momento in poi, salvo diversa indicazione, tutti i brani tratti da testi stranieri sono da ritenere frutto di nostra traduzione in italiano.

[6] A riprova del fatto che il concetto di misurazione, così come concepito nel campo delle scienze fisiche, è largamente accettato e adottato dal senso comune come quello prevalente, si noti quanto questa definizione di Michell sia sovrapponibile a quella del vocabolario Treccani riportata in precedenza.

[7] Le proprietà operativizzabili attraverso conteggio sono definite cardinali discrete da Marradi [es. 2007].

numero infinito di stati impercettibilmente differenti[8]. Più in generale, alle spalle di questa concezione classica della teoria della misurazione, Michell individua quello che chiama il *paradigma euclideo*, fondato sulla definizione (attribuita per l'appunto a Euclide) dell'uguaglianza dei rapporti tra grandezze geometriche[9]. In tale ottica, cui fu debitrice la definizione stessa dei numeri reali data da Dedekind a fine Ottocento, si dimostra che i "rapporti tra grandezze di una quantità misurabile sono isomorfici ai numeri reali positivi e la misurazione è la stima di tali rapporti" [Michell 2007, 99]. Una stima sempre possibile in via di principio, poiché esiste sempre un numero reale che rappresenti il rapporto tra un *misurando* e un'unità di misura.

Il cuore del paradigma euclideo, per ciò che riguarda la nostra discussione, è che qualsiasi grandezza (ossia, *specifica quantità continua*) di un certo tipo può essere *descritta* come multiplo di una grandezza standard, l'*unità di misura*. Nel caso di grandezze cosiddette *estensive* le operazioni di accertamento avvengono attraverso una procedura di *concatenazione* (e in questo senso Helmholtz [1887] le definisce "additive"). Esistono anche proprietà qualitative che sono considerate grandezze *intensive*, poiché tecnicamente la determinazione del rapporto con l'unità di misura non avviene attraverso la concatenazione di quest'ultima, ma la cui natura additiva può essere dimostrata in altri modi. Ciononostante, anche queste sono costituite da *continua* misurabili, soggiacenti alla logica del paradigma euclideo della misurazione.

Seguendo il filo di questo ragionamento, il presupposto di qualsiasi pretesa di misurazione (a rigore, anche di costrutti mentali) risiede nel poter definire proprietà quantitative continue. Su questo spartiacque, secondo Michell, si annida la difficoltà delle scienze umane: esse non sarebbero mai riuscite – sostiene – a dimostrare la natura quantitativa e continua delle proprietà attitudinali che taluni reclamano di misurare. Le evidenze empiriche mostrerebbero – prosegue – che esse sono caratterizzate, al limite, solo da stati ordinabili. Inferire, tuttavia, dall'esistenza di un *ordine* quella di una *struttura quantitativa* è esattamente ciò che egli chiama la "fallacia psicometrica", fondata sull'argomento (espresso per esempio da Bergson [1889]) che l'esistenza di un ordine

[8]Tali proprietà possono essere operativizzate in variabili cardinali continue [ibid.]

[9] Il filosofo greco mostra che il rapporto tra due segmenti x e y è uguale al rapporto tra altri due segmenti w e z se e solo se, scelta qualsiasi coppia di numeri interi a e b, si verifica che i) $ax < by$ se e solo se $aw < bz$; ii) $ax = by$ se e solo se $aw = bz$; iii) $ax > by$ se e solo se $aw > bz$. In virtù di queste relazioni, qualsiasi rapporto tra grandezze si situa quindi lungo la serie dei numeri reali. Sul tema si veda anche Moretto [2012].

tra oggetti rispetto a una proprietà posseduta, e quindi una relazione di "maggiore o minore" rispetto a essa, implichi logicamente l'esistenza di una *distanza* tra tali oggetti e quindi la possibilità logica, in ultima istanza, di quantificarla. Priva di adeguati "puntelli" empirici e minata dalla fallacia psicometrica, qualsiasi pretesa di misurazione nelle scienze umane e sociali sarebbe quindi, dal punto di vista di Michell, infondata, spinta dal bisogno di reclamare in modo fittizio uno *status* scientifico pari a quello delle scienze naturali (definite spesso *hard*, in contrapposizione a quelle umane, più deboli e quindi etichettate come *soft*).

Dello stesso avviso, nel campo della ricerca sociale, è anche Marradi [2007], il quale adotta una definizione della misurazione che ben aderisce a quella classica di stampo euclideo. Egli riconosce al processo di misurazione una serie di prerequisiti: una proprietà da misurare definibile lungo una dimensione continua; la stipulazione di un'unità di misura universalmente riconosciuta[10]; il confronto (e con ciò il rapporto) tra la grandezza da misurare e l'unità di misura opportunamente ripetuta. Partendo da questi elementi, Marradi rileva alcuni aspetti critici per la misurazione degli stati mentali delle persone. Solleva anche lui osservazioni simili a quelle di Michell sul carattere solo ipotetico della natura continua e quantitativa delle proprietà in questione, essendo non direttamente rilevabili. Ciò comporta la problematicità di definire un'unità di misura. In particolare, però, pone l'accento su un ulteriore aspetto. Infatti, argomenta Marradi, quand'anche fosse legittimo immaginare e definire un costrutto latente mentale come una dimensione continua e individuare su di essa intervalli equidistanti, rispetto alle scienze fisiche cambierebbe comunque, radicalmente, la meccanica della misurazione. Il fatto è che non siamo mai alla presenza di un confronto diretto ed esterno tra lo strumento di rilevazione e la grandezza in questione, ma è il soggetto stesso a valutare la propria posizione lungo il *continuum*, sia direttamente (collocandosi su una scala) sia indirettamente (attraverso una batteria d'indicatori). Ciò rappresenta una differenza epistemologica sostanziale con la misurazione delle dimensioni fisiche, un limite tecnicamente insormontabile data l'impossibilità di osservare direttamente i processi mentali in

[10] Il fatto che sia universalmente riconosciuta rispecchia il processo, avvenuto nelle scienze fisiche, di costruzione di un sistema internazionalmente condiviso di unità di misura. In linea teorica, la logica del ragionamento di Marradi sembrerebbe richiedere piuttosto che l'unità di misura sia quantomeno definibile in modo tale da poter essere compresa intersoggettivamente e riprodotta in occasioni diverse.

un individuo e di controllare che i meccanismi attivati siano i medesimi tra individui diversi. Marradi, come Michell, ritiene quindi che manchino i presupposti affinché le scienze sociali possano reclamare la pretesa di misurare *strictu sensu* dimensioni attitudinali, valoriali ecc. Se possibile, porta ancora più a fondo la sua critica, spostandola dal piano empirico della mancanza di evidenze riguardo alla natura quantitativa degli atteggiamenti, a quello logico della natura stessa della procedura di rilevazione. L'unico stratagemma, si sostiene, per rivendicare una qualche forma di misurazione è di allargarne il significato fino a comprendere sotto la sua etichetta qualsiasi operazione di rilevazione e classificazione[11].

1.2 La misurazione come rappresentazione

1.2.1 Dalla logica di Russell alla teoria di N.R. Campbell

Un contributo classico è quello del fisico ed epistemologo Norman Robert Campbell [1920], che definisce la misurazione come l'assegnazione di numeri[12] per rappresentare proprietà in accordo con leggi scientifiche. Alla base di questa accezione troviamo alcuni elementi sviluppati nella teoria rappresentazionale del logico Russell [1903], per il quale la misurazione è un isomorfismo tra due serie, ossia una corrispondenza biunivoca tra due insiemi: una classe di proprietà dello stesso tipo, ordinate in base alla loro grandezza (che formano un *attributo quantitativo*) e i numeri di un certo tipo (la serie dei numeri interi, razionali, reali ecc.). Per Russell condizione necessaria per parlare di misurazione è che esista una tale corrispondenza isomorfica. Egli assegna importanza principalmente pratica, più che logica, al *come* stabilire, scegliere tale corri-

[11] E' quello che esplicitamente fanno Nunnally e Bernstein [1994], per esempio, definendo la *classificazione* come una specie del genere *misurazione*. In particolare, con riferimento alla teoria delle scale di Stevens (di cui si discuterà successivamente), la classificazione sarebbe la forma di misurazione nel caso delle scale nominali, mentre negli altri casi si parlerebbe di *scaling*. Per cui, anche chi compie una classificazione starebbe misurando. Chi scrive ritiene sia più rigoroso, sulla scia di Marradi, definire piuttosto la misurazione come una forma particolare di classificazione, trattandosi in ogni modo dell'assegnazione di un caso a una *classe* (nella fattispecie della misurazione classica, a un intervallo definito dalla sensibilità dello strumento).

[12] Nell'originale inglese il termine è *numerals*, cioè i numeri intesi come cifre, come significanti, distinto da *numbers*, cioè i numeri veri e propri come valori.

spondenza: è ragionevole che lo scienziato desideri trovare un isomorfismo che rispecchi, per esempio, le relazioni interne alle due serie (le grandezze e i numeri). Egli riconosce, inoltre, che vi sono delle grandezze connotate non solo da relazioni d'ordine, ma anche additive, che possiedono cioè le caratteristiche dei numeri. Per il *logico* Russell, comunque, ciò è appunto una questione di ordine pratico, più che di rilevanza teorica.

La definizione di Campbell poggia l'impianto rappresentazionale su un piano filosofico realista, dovuto anche al retroterra scientifico dell'autore. Per lui il problema empirico ha una rilevanza centrale, è discriminante per stabilire un criterio di misurabilità. Assegnare numeri è la procedura per misurare, ma per parlare di misurazione è necessario che ciò sia fatto in accordo con leggi scientifiche, cioè che i rapporti numerici siano intrinseci agli attributi misurati. Le *quantità* (o anche *grandezze del tipo A*) sono attributi omomorfici rispetto ai valori numerici: ne possiedono il carattere di additività, empiricamente dimostrabile attraverso procedure di concatenazione. Non è solo una classe di proprietà *ordinabili* in base alla grandezza. La misurazione di una grandezza del tipo A è chiamata *fondamentale* da Campbell. L'unica altra forma di misurazione presa in considerazione è quella *derivata*, che riguarda le cosiddette *qualità* o *grandezze del tipo B*: si tratta di attributi coinvolti in leggi fisiche di natura, come la densità, e quindi accertabili come rapporti o costanti tra proprietà misurate in modo fondamentale[13]. Nella riflessione di Campbell non trova spazio una posizione per le tecniche di rilevazione delle scienze psicologiche e sociali.

Mezzo secolo più tardi, lo psicologo Nunnally [1978] utilizza una definizione decisamente congruente con quella precedente: la misurazione consiste in *regole* per l'assegnazione di *numeri*[14] a oggetti in modo tale da *rappresentare quantità* di attributi. Anche qui si parla di proprietà; si afferma la necessità di stabilire regole di corrispondenza per l'assegnazione dei numeri, ossia una funzione di rappresentazione che rifletta una struttura empirica su una numerica. In questo caso, però, le dimensioni psicologiche sono considerate dall'autore proprietà misurabili, diversamente da Campbell. La chiave di volta per comprendere tale mutamento risiede nella riflessione teorica di S. Smith Stevens, che ha avuto un impatto fortissimo nel campo delle discipline psicologiche e

[13] La distinzione tra quantità e qualità (grandezze di tipo A e di tipo B) richiama quella citata tra grandezze estensive e intensive.

[14] Si noti che Nunnally usa il termine *numbers* nella sua definizione.

sociologiche, tanto da diventarne quasi un paradigma, generalmente accettato da chiunque rivendichi un'attività di misurazione. La sua diffusione è stata pari al dibattito critico che ha sollevato da più parti.

1.2.2 L'attacco alla misurazione in psicofisica e la proposta di Stevens

Il contributo di Stevens parte dal suo interesse di studio, la misurazione dell'intensità delle sensazioni, e si sviluppa dalla seconda metà degli anni '30. Siamo nel campo della ricerca psicofisica, tradizione inaugurata quasi un secolo prima da pionieri della psicologia come Weber, Fechner, Wundt. Alla fine degli anni '30 le discipline psicologiche subiscono una dura messa in discussione delle proprie procedure di misurazione. L'Associazione Britannica per l'Avanzamento della Scienza istituisce una commissione per discutere dello *status* della misurazione in psicologia [Michell 2007; Pitrone 2012], tra i cui membri compare lo stesso Campbell. La commissione parte dal presupposto che gli psicologi, come tutti gli altri ricercatori, debbano dimostrare che le proprie scale sono in grado di misurare, e farlo attraverso prove empiriche.

La conclusione, in quel momento, fu che gli psicologi (e gli studiosi sociali in generale) non fossero capaci di giustificare i propri processi di rilevazione alla luce del concetto all'epoca corrente di "misurazione", incardinato come visto in un'ottica rappresentazionale e realista (interna a quanto definito come *paradigma euclideo*) secondo cui una corrispondenza numerica è dettata da una regola di rappresentazione che *riproduce* una struttura effettivamente presente nella realtà[15]. Nel rapporto finale si sottolinea che *misurazione*

> è semplicemente una parola utilizzata convenzionalmente per denotare certe idee. Usarla per denotare altre idee non allarga il suo significato ma lo distrugge" [Ferguson *et al.* 1940].

[15] Nel rapporto finale della commissione si legge: "Qualsiasi legge che si proponga di esprimere una relazione quantitativa tra l'intensità della sensazione e dello stimolo non è semplicemente falsa ma è, di fatto, priva di significato salvo che e fintanto che non possa essere dato un significato al concetto di addizione applicato alla sensazione." [Ferguson *et al.* 1940, 245].

E', nella sostanza, la stessa critica in seguito ripresa, come visto, da Marradi. Il quale aggiunge tra l'altro che nello studio delle dimensioni latenti l'impresa è ancora più difficile rispetto al campo delle sensazioni e delle percezioni, che perlomeno hanno una base fisica che può servire da riscontro [Marradi 2007].

Stevens [1946, 677] sostiene che "la misurazione è definita, nel senso più ampio, come l'assegnazione di numeri[16] a oggetti o eventi sulla base di regole[17]". Si osservi come, a differenza della definizione di Campbell e di quella successiva di Nunnally, in quella di Stevens non si faccia riferimento agli attributi: la misurazione riguarda gli oggetti e non le loro proprietà. In seconda battuta, il compito riconosciuto da Stevens alla misurazione è di "assegnare numeri" agli oggetti, laddove nella concezione classica delle scienze fisiche si tratta di *scoprire* le relazioni esistenti intrinsecamente tra proprietà sotto forma di *rapporti numerici* [Michell 2007]. In questo modo si separano, almeno teoricamente, le procedure di costruzione dei modelli di misurazione dal riconoscimento di un contenuto empirico intrinseco alla misura e si sposta l'attenzione sulla struttura del modello stesso. Infine, la lettera della definizione di Stevens sembra comportare una liberalizzazione integrale del campo di misurabilità: ogni oggetto cui si possa immaginare di assegnare un numero, una cifra, un'etichetta numerica, è di per se stesso misurabile.

I critici di Stevens vedono in questo passaggio quell'ampliamento del significato della misurazione che finirebbe per distruggerlo, per usare le parole della relazione della commissione Ferguson. Resta che questo impianto ha fornito agli scienziati sociali una teoria della misurazione su cui incardinare quasi tutte le riflessioni e gli sforzi successivi in ambito psico- e sociometrico. Per certi aspetti, siamo di fronte a un ribaltamento dell'impostazione realista di Campbell e di quella più "liberale" elaborata in seguito da Cohen e Nagel [1934]. Sicuramente per Stevens non si tratta solo di trovare stratagemmi definitori per allargare il campo di legittimità dei processi di misurazione. Dietro la sua riflessione possiamo rintracciare alcuni segni dello spirito del suo tempo.

L'*operazionalismo* di Bridgman [1927] da un lato (in sintesi, l'idea che un significato univoco di un concetto derivi dalle operazioni svolte per definirlo) e l'interesse

[16] Nell'originale dell'autore si parla di *numerals*.

[17] Scrive ancora Stevens, qualche pagina dopo: "Alla commissione britannica, quindi, possiamo azzardarci a suggerire a titolo di conclusione che la definizione di più ampie vedute e utile di misurazione è, come ha consigliato uno dei suoi membri, "l'assegnazione di numeri alle cose in modo da rappresentare fatti e convenzioni su di essi." [ibid., 680]

dei *neopositivisti* per le strutture logico-formali, il linguaggio della matematica e i sistemi numerici, aiutano a capire l'impostazione di Stevens riguardo ai sistemi di misurazione, caratterizzata da un maggiore interesse per la struttura interna delle scale e per la loro logica di costruzione: il rapporto con la realtà e con la dimensione empirica, seppure importante, è un passaggio successivo, un'applicazione posteriore nel tempo, incidentale rispetto alla coerenza delle scale e alle loro regole d'invarianza.

1.2.3 La teoria delle scale di Stevens

A questo proposito, Stevens elabora un contributo importantissimo, largamente adottato in tutte le scienze sociali: la distinzione tra le possibili funzioni di rappresentazione, ossia la classificazione dei tipi di scale di misurazione. Il criterio che utilizza, come accennato, è quello delle operazioni di manipolazione numerica ammissibili su una scala in modo da garantire l'invarianza rispetto alle regole dell'assegnazione. In sostanza, la domanda da porsi è "quali operazioni posso applicare ai valori della mia scala senza alterarne il significato e la struttura interna?". Egli, infatti, sostiene che

> la misurazione esiste in una varietà di forme e le scale di misurazione ricadono all'interno di certe classi definite. Queste classi sono determinate sia dalle operazioni empiriche invocate nel processo di "misurazione", sia dalle proprietà formali (matematiche) delle scale. Inoltre (…) la manipolazione statistica che può legittimamente essere applicata ai dati empirici dipende dal tipo di scala sulla quale sono ordinati i dati. [1946, 677]

Stevens individua, com'è noto, quattro tipi di scale, ossia livelli di misurazione: nominale, ordinale, intervallare, di rapporto.

Una *scala nominale* [18] (o dicotomica, o categoriale) "permette d'individuare nell'oggetto d'indagine solo la presenza (1) o l'assenza (0) di una data proprietà: quindi,

[18] Secondo Marradi [1980] tale etichetta è una *contradictio in adjectu*: il concetto di *scala* introduce l'idea di una successione crescente o decrescente, di una serie di gradini; invece il concetto di *nome* riporta a una distinzione tra classi solo qualitativamente distinte, non necessariamente collocabile lungo una dimensione ordinabile.

l'equivalenza o meno di due oggetti rispetto a quella proprietà" [Bruschi 1999, 82]. E' invariante rispetto a qualsiasi trasformazione uno-a-uno, che sostituisca i valori della scala mantenendo il principio di equivalenza nel processo di assegnazione. Secondo Stevens [1946, 678] in questo caso "i numeri sono usati solo come etichette o simboli", al pari di qualsiasi altro segno che potrebbe essere utilizzato. Se ci attestiamo sulla convenzione che "misurare" significhi assegnare numeri in conformità a regole, allora anche questa è misurazione.

Una *scala ordinale* permette di distinguere se "una proprietà è presente in un oggetto più che in un altro" [Bruschi 1999, 82], senza dare informazioni quantitative sull'entità della differenza[19]. La struttura di una scala ordinale rimane invariante rispetto a qualsiasi trasformazione monotonica, che cioè preservi l'ordine dei valori rispetto alla proprietà.

Una scala *a intervalli*, oltre a istituire un ordinamento sulla base di una proprietà, stabilisce anche una suddivisione della proprietà in una successione d'intervalli uguali, su cui collocare gli oggetti. Il punto zero non ha una valenza fissa, ma è stabilito *arbitrariamente*. In queste scale è possibile determinare la distanza tra due oggetti e confrontare le differenze tra le distanze degli intervalli. Una scala intervallare è invariante rispetto alle trasformazioni lineari, cioè se tutti i valori della scala sono moltiplicati per un fattore positivo e/o sommati a una costante (positiva o negativa)[20].

Una *scala di rapporto* è tale se, oltre all'ordinalità e alla presenza d'intervalli uguali, individua anche un *punto di zero assoluto*, cioè di assenza della proprietà. Una scala di rapporto è invariante rispetto alle trasformazioni proporzionali, cioè la moltiplicazione di ogni valore per una costante positiva[21].

Anche la classificazione di Stevens può essere vista come una scala di tipo ordinale, rispetto alla capacità informativa e alla rigidità della struttura sintattica: nel passare dalle scale nominali a quelle di rapporto, aumenta la capacità informativa [Bruschi,

[19] Marradi [1981] fa notare che dietro l'etichetta di "misurazione ordinale", così come già definita da Campbell, convivono principalmente due diverse operazioni: l'assegnazione dei soggetti a categorie ordinate e il confronto tra le quantità di una proprietà possedute da più soggetti. In entrambi i casi egli afferma che il termine "misurazione" è alquanto ardito: nel primo caso perché non si sta misurando, bensì assegnando soggetti a classi; nel secondo caso perché nelle scienze sociali quasi mai esiste un'unità di misura per tale confronto e sarebbe quindi forse più cauto parlare di "confronto quantitativo".

[20] Esempi di scale a intervalli sono quelle Celsius e Fahrenheit per la temperatura.

[21] Le scale di rapporto sono la norma nelle scienze fisiche [Coombs 1950].

1999], mentre diminuiscono le trasformazioni legittime per garantire l'invarianza (quindi aumenta la rigidità sintattica).

Generalmente, i fautori della misurazione nelle scienze psico-sociali sostengono che il livello di misurazione intervallare sia quello raggiunto dalle scale abitualmente utilizzate[22], mentre in pratica non si riscontrano proprietà con uno zero assoluto teorico su cui applicare scale di rapporto. Per i critici, invece, resta immutata la validità dei giudizi espressi a suo tempo dalla Commissione Ferguson: non ci sarebbero prove sufficienti a sostenere che le proprietà mentali latenti possano ragionevolmente andare oltre il livello degli stati ordinabili.

1.3 Rapporto tra strutture e funzioni e il ruolo degli assunti

Michell [2007] rileva un aspetto teorico importante dell'approccio di Stevens, che risiede nella possibilità che le regole di assegnazione, così come descritte da quest'ultimo, possano essere interpretate alla stregua di *funzioni che generano strutture*, specularmente all'ipotesi realista. Difatti, la determinazione delle relazioni strutturali sulle varie scale[23], in un'ottica operazionalista, è definita dalle operazioni di assegnazione stesse e quindi non necessariamente da un rapporto con una struttura empirica indipendente e data, qualsiasi sia lo strumento di rilevazione utilizzato. Questa interpretazione è avvalorata da alcune osservazioni. Nella definizione di misurazione riportata poco fa (vedi nota 17), Stevens precisa che l'assegnazione di numeri a oggetti deve avvenire attraverso regole, così da rappresentare "fatti o *convenzioni* su di essi" (corsivo nostro). Una scala, quindi, può rappresentare non solo una struttura empirica, qualcosa

[22] Si rimanda alla Parte Terza, dove sarà affrontato l'argomento e riportata una serie qualificata di citazioni in questo senso. A titolo d'esempio, emblematico di questa posizione con riferimento alla misurazione in psicologia e, nello specifico, al modello di Rasch (che sarà il centro della nostra attenzione), scrivono Robusto e Anselmi [2012, 49]: "Si è visto come anche in psicologia sia possibile introdurre una misurazione che abbia le stesse caratteristiche fondamentali della misurazione ottenuta in fisica. (…) Il modello [di Rasch] permette di definire una cornice teorica di riferimento entro la quale realizzare anche in psicologia un sapere rigoroso dal punto di vista misurativo, analogamente a quanto già avviene in tutte le altre scienze."

[23] In tutte le scale si determinano relazioni di equivalenza; relazioni d'ordine si determinano in tutte le scale, tranne quelle nominali; relazioni di uguaglianza d'intervalli (equivalenza di differenze) si determinano nelle scale intervallari e di rapporto; solo in queste ultime si determinano relazioni di proporzionalità (equivalenza di rapporti).

di esistente in natura, *in re* a prescindere dall'attività di rilevazione umana: una funzione di rappresentazione può ricostruire convinzioni raggiunte tramite un accordo tra esseri umani (in questo caso scienziati), convenzioni, elementi giustificati magari da considerazioni di ordine teorico-ipotetico. Il modello di misurazione può, in ultima istanza, andare a definire cosa dobbiamo misurare. Ed è questo, ci pare, il senso ribadito da Stevens laddove esplicitamente sostiene [1968] che i numeri che emergono dalle misurazioni si portano dietro l'impronta delle operazioni con cui sono stati ottenuti e che è compito del ricercatore decidere il loro significato e la loro importanza relativamente all'avanzamento della ricerca empirica.

Quello del rapporto tra struttura del *misurando* e funzione di rappresentazione è un punto epistemologico molto importante, di cui può essere utile riassumere i termini. Se, in conformità a considerazioni empiriche e teoriche, si ritiene che una scala di misurazione debba essere il riflesso di rapporti numerici tra proprietà esistenti nella realtà data, allora sarà coerente pensare che sia la struttura a generare la funzione di rappresentazione, vincolando così le regole di assegnazione. Diversamente, se l'accento è posto sulle regole con cui determinare le proprietà della scala, attraverso le quali interpretare quelle della realtà, si parlerà di funzioni che generano la struttura. Ciò non significa che non debba esserci un rapporto di non contraddittorietà tra la struttura empirica e quella formale, testabile nell'ambito della riflessione sulla validità, ma identifica un approccio diverso rispetto a quello classico.

Al di là delle dispute e delle posizioni epistemologiche, l'orientamento espresso da Stevens pare effettivamente confarsi alla condizione in cui avviene nella pratica la rilevazione di dimensioni attitudinali nel campo delle scienze umane. La natura latente di tali oggetti di studio, già richiamata, e soprattutto l'impossibilità, almeno pratica, di produrre un'attività sperimentale di manipolazione delle variabili comparabile con quella delle scienze fisiche[24], assegna, di fatto, un ruolo decisivo a tutta una serie di assunti

[24] Questo significa l'impossibilità di controllare in modo indiretto il rapporto dei costrutti "non visibili" con quelli direttamente manipolabili, tenendo invariati tutti gli altri fattori considerati come influenti. Studiosi come Trendler [2009] avanzano una posizione radicalmente più pessimistica di quella di Michell: la misurazione nelle scienze umane sarebbe impossibile in via di principio. Sulla scia della posizione classica di Mill, egli sostiene che poiché la mente umana è influenzata da una serie indefinita di fattori esterni e di *disturbi sistematici* (nel senso sperimentale e non clinico del termine) e poiché questi non possono essere controllati e manipolati da un apparato sperimentale, nessun tipo di esperimento può essere

teorici (necessari a costruire e sostenere i modelli di misurazione) diversi da caso a caso. Giova, ovviamente, ripetere che i modelli devono essere supportati da strumenti operativi, i cui risultati possono o no contraddire gli assunti di base. Cionondimeno, la dimensione teorica *a priori* svolge un ruolo fondamentale nel plasmare le ipotesi di funzionamento dei meccanismi sottostanti i fenomeni studiati e nel selezionare le dimensioni considerate rilevanti. Se nelle scienze sociali si può parlare di misurazione, la teoria di Stevens sembra predisporre il perimetro entro cui essa può essere giustificata.

Dall'altra parte, come mette in luce la sintetica ricostruzione compiuta, sotto l'etichetta di "misurazione" convivono significati diversi, a volte anche tra loro lontani, spesso distanti dal senso originario del paradigma euclideo. Bruschi [1999, 80-1] riassume la questione, adottando, da un lato, il termine "misurazione" per certe procedure di rilevazione delle scienze sociali, ma precisando, dall'altro, che nella pratica di quest'ultime il significato di quella parola si riferisce in generale alla "rappresentazione di una struttura empirica su una struttura formale" secondo una corrispondenza *morfica*. Perciò è necessario tenere ferma la consapevolezza che, rispetto all'accezione più ristretta della misurazione nelle scienze fisiche, quella della ricerca sociale "denota qualsiasi processo di assegnazione di valori formali e (…) viene utilizzata come sinonimo di rappresentazione". Inoltre, l'autore riprende la distinzione tra grandezze estensive e intensive: mentre nelle scienze naturali si hanno strumenti fisici con cui poter determinare la misura, nelle scienze sociali presiedono criteri diversi. Questo rappresenta, a nostro modo di vedere, una discontinuità epistemologica forte tra i due tipi di misurazione. Bruschi pone l'accento sul fatto che la concatenazione additiva e il confronto tra grandezze delle scienze fisiche sono sostituite dai criteri di frequenza, di manifestazione e di assegnazione soggettiva delle scienze sociali. La "misurazione" avviene in un quadro tale che, se da un lato "possiamo dare per scontata l'"esistenza" delle strutture formali (sono quelle del patrimonio cognitivo della matematica) le strutture empiriche corrispondenti devono essere *costruite*" [ibidem, 90]. Soprattutto nel caso di strutture intervallari e di rapporto, non direttamente rilevabili in termini empirici in ambito sociale,

immaginato in questo ambito e quindi non avrebbe logicamente significato parlare di una misurazione in senso classico (che l'autore definisce "galileiana"): se i fenomeni psicologici non sono dipendenti o non possono essere resi dipendenti da un insieme maneggevole di condizioni, allora non sono misurabili. Il fallimento nel controllo di disturbi sistematici renderebbe quindi impossibile la scoperta di una struttura quantitativa.

“ai problemi di rappresentazione si aggiungono quelli di operazionalizzazione” [ibidem].

Nel corso del testo incontreremo altri aspetti riguardanti la teoria della misurazione, che saranno al centro della riflessione. Quanto riassunto sinora, ben lungi dall’essere esaustivo, aiuta tuttavia a inquadrare alcuni punti principali. Nelle scienze fisiche è abbastanza pacifico cosa significhi misurare e quali siano i requisiti di una misura corretta. Vi è anche una certa chiarezza su cosa differenzi una misurazione fondamentale (o additiva) da una derivata. Una lunga tradizione, che qualcuno ricomprende sotto l’etichetta di “paradigma euclideo”, e la natura degli oggetti di studio, permettono di definire in modo condiviso concetti come “quantità”, “concatenazione”, “additività” ecc.

I problemi epistemologici sorgono quando s’importa tale apparato concettuale dentro il campo di studi delle scienze umane. La natura non osservabile e latente di oggetti cognitivi come atteggiamenti, opinioni, dimensioni valoriali pone alcuni limiti, che possono essere superati in modo ipotetico introducendo assunti teorici provvisori laddove manchino riscontri e corroborazioni empirici. La prima questione che abbiamo osservato riguarda il grado in cui la definizione della misurazione, così com’è stata elaborata in campo fisico, possa adattarsi allo studio dell’uomo e quali modelli rispettino un arco di requisiti tale da renderli confrontabili con gli standard della misurazione classica. Uno degli scopi della presente ricerca è esercitare questa riflessione sulla fattispecie del modello di Rasch.

La seconda questione emersa riguarda fino a che punto la definizione di misurazione possa assumere un significato più ampio di quello originale senza essere snaturata e perdere la sua legittimità e quando, invece, essa mantenga del concetto originario solo il nome. Ciò riguarda il rapporto epistemologico tra le procedure di misurazione e le funzioni di assegnazione, da una parte, e la struttura empirica dall’altra. Questo si riflette anche, di conseguenza, sul tema assolutamente strategico del livello di misurazione che si reputa legittimo rivendicare, poiché da questo discendono a catena tutte le condizioni di possibilità relative, per esempio, alle tecniche di analisi utilizzabili legittimamente, agli indici, alle misure di posizione ecc.

A questo punto è possibile introdurre brevemente alcuni concetti basilari, tratti dalla sterminata letteratura sul tema dello *scaling* nelle scienze umane e sociali, dedi-

cando poi qualche attenzione ad alcuni modelli classici, particolarmente indicativi per il seguito del testo.

2 Quantificare atteggiamenti e opinioni: cenni introduttivi al concetto di *scaling*

Nunnally [1978, 35] definisce un modello di *scaling* come un "piano internamente coerente per lo sviluppo di una nuova misura". La teoria della misurazione nelle scienze sociali è intrecciata intimamente con il concetto di *scaling*: stabilire cosa significhi misurare conduce a definire cosa sia una scala di misurazione e viceversa. Abbiamo però ritenuto di presentare in successione, e non insieme, i due argomenti perché, come ricorda Marradi [1981, 607], "misurazione" e "scala" non sono sinonimi e, anzi

> i due termini, con i relativi concetti, dovrebbero essere tenuti accuratamente distinti, e non solo per ragioni logico-formali. Un discorso centrato sulla misurazione, infatti, dirige l'attenzione su una sequenza di operazioni intellettuali e pratiche, con le relative caratteristiche, condizioni e limiti. Un discorso centrato sulle scale, invece, dirige l'attenzione sulle proprietà logico-matematiche del prodotto finale di tali operazioni. Negli anni '20 (…) Norman Campbell aveva (…) richiamato l'attenzione sulle procedure, influenzando il lavoro di Thurstone e della sua scuola; ma alla fine degli anni '40, i (…) lavori di Stevens invertirono la tendenza, portando alla ribalta le scale e le loro proprietà matematiche.

Per chi abbraccia la proposta teorica di Stevens sulla misurazione e la sua classificazione delle scale, è coerente sostenere [Bruschi 1999, 82] che "nelle scienze sociali la rappresentazione degli stati delle proprietà è espressa dalle scale". In questo modo, il termine "scala" diventa sinonimo di quello "variabile". Ciò è giustificato dal fatto che Stevens, nella sua teoria, trova posto per le scale nominali, le cui categorie non sono connesse da rapporti ordinali o quantitativi. Tutti i tipi di variabili costituiscono, in questo modo, scale.

Su quest'ultimo aspetto, come già notato (vedi nota 11), Nunnally e Bernstein sollevano un'eccezione, parlando di "classificazione" nel caso di variabili nominali e riservando il concetto di "scala" alle variabili che rispecchiano almeno relazioni d'ordine. Marradi (vedi nota 18) è ancora più netto, rilevando una forte incongruenza nell'utilizzo di un termine come quello di "scala", che suggerisce l'idea di diversi livelli, per variabili che non esprimono proprietà che aumentano e diminuiscono.

Cannavò [2003], illustrando la definizione di un dizionario della lingua inglese, scrive che il concetto di *scaling* rappresenta "esattamente la concezione estesa di misurazione sociale", poiché significa "un processo inteso a graduare (*range*) e a trovare la misura delle cose (*the size of things*) ma (…) '*by comparing them, measuring them or counting them*'". Questa definizione, pur rifiutando un'accezione di scala aderente alla versione fisicista della misurazione, chiarisce l'idea che la costruzione di scale di misurazione abbia in qualche modo a che fare con la "dimensione delle cose": le tre parole chiave sono "comparare, misurare e contare", che potrebbero richiamare, per certi aspetti, i ruoli delle variabili ordinali, di quelle cardinali continue e di quelle cardinali discrete. Riteniamo utile questa definizione nel caso si vogliano individuare quei modelli che nella pratica delle indagini hanno in comune la ricerca di dimensioni in vario modo quantitative.

2.1 Alcuni criteri di classificazione dei modelli di *scaling*

Tra gli altri, McIver e Carmines [1981] individuano una serie di criteri sulla cui base è possibile distinguere i modelli di *scaling*[25].

Innanzitutto, essi possono essere utilizzati per scalare separatamente i soggetti o gli *item*, ovvero entrambi contemporaneamente[26]. I soggetti sono, solitamente, le perso-

[25] Si tenga presente che diversi autori, esplicitamente o meno, distinguono opportunamente tra *modello di scaling* (che riguarda l'approccio più generale e l'insieme di assunti che compongono il quadro teorico per la costruzione di una scala) e la *tecnica di scaling* (che identifica una specifica procedura che, aderendo ai principi di un modello di *scaling*, prevede le operazioni necessarie alla costruzione della scala e all'assegnazione dei punteggi). Si veda più estesamente sul tema, ad esempio, Maggino [2004a; 2004b; 2007].

[26] Un esempio di *scaling* orientato al soggetto è quello di Likert; esempi di *scaling* orientati agli *item* sono quelli elaborati da Thurstone; *scaling* orientati alla stima simultanea di soggetti e *item* sono quello

ne i cui atteggiamenti o le cui opinioni si è interessati a collocare lungo la dimensione latente studiata. Gli *item* sono invece ciò che, nel lessico d'ispirazione comportamentista, chiameremmo "stimoli": sono i mezzi attraverso i quali si manifesta la proprietà in questione. In un testo classico sul tema, Torgerson [1958] afferma che essi sono "qualsiasi cosa che può servire a dividere i soggetti in due o più categorie mutuamente esclusive ed esaustive (…) una tipica domanda di questionario concernente un atteggiamento o un'opinione, informazioni biografiche o credenze (…) comportamenti manifesti (…) qualsiasi cosa risulti in una classificazione dei soggetti in categorie, con la restrizione che ciascun soggetto deve essere in una e una sola categoria dell'*item*". Il rapporto tra soggetti e *item* richiama, in un certo senso, quello tra l'oggetto fisico e lo strumento di misurazione: la persona possiede un certo grado di una proprietà (diciamo, un atteggiamento), come l'oggetto fisico è caratterizzato da un certo attributo; le affermazioni, le frasi, gli stimoli che costituiscono gli *item* sono l'espressione della varietà di componenti che nell'insieme costituiscono la dimensione della proprietà, così come lo strumento di misurazione dell'oggetto fisico è una manifestazione dello stesso attributo, di cui rappresenta l'unità di misura (o suoi multipli o sottomultipli). Tale parallelo richiede ovviamente una lettura *cum grano salis*, alla luce delle considerazioni già introdotte in precedenza e del fatto che "la costruzione dello strumento per la misurazione intensiva si basa su operazioni spesso implicite, carenti sul piano della pubblicità e dell'intersoggettività" [Bruschi 1999, 95].

Un secondo criterio è l'eventuale forma funzionale che descrive la relazione tra l'attributo espresso dall'*item* e la probabilità del soggetto di appartenere a ciascuna delle possibili categorie di risposta. Da tale rapporto si può capire se il modello è deterministico o stocastico, se è cumulativo oppure descrive una relazione simmetrica, se la propensione all'accettazione dell'*item* cresce o decresce all'aumentare dell'intensità dell'attributo ecc.

Il terzo criterio riprende l'importante classificazione, proposta da Coombs [1964], delle procedure per determinare la base empirica del modello. Nella prima classe troviamo i procedimenti di *scelta preferenziale*, che richiedono l'ordinamento degli stimoli lungo una dimensione sottostante, come quando si chiede a un intervistato di classifica-

dello scalogramma di Guttman e lo stesso modello logistico di Rasch (dei quali si tratterà successivamente).

re una serie di personaggi politici lungo una scala di preferenza o di vicinanza alle proprie posizioni. La seconda classe riguarda le risposte a *stimolo singolo*, ottenute chiedendo ai soggetti di giudicare, valutare, accettare ecc. ogni *item* separatamente, uno per volta, come accade nella tecnica di Likert, in cui ognuno esprime il proprio grado di accordo rispetto a ciascuna affermazione di una batteria. La terza classe comprende la *comparazione degli stimoli*, definita dalla scelta di un solo *item* (tra un insieme di altri presentati simultaneamente) secondo un certo criterio, come quando si indica la città in cui si preferirebbe vivere tra quelle presentate in una lista di capoluoghi di regione italiani. L'ultima procedura concerne il criterio di *somiglianza*, per il quale si procede presentando ai soggetti tutte le possibili coppie di *item* a due a due, chiedendo loro quale sia quella con i due *item* più simili e ricostruendo così in seguito gli attributi alla base della scelta.

Il quarto criterio è quello della n-dimensionalità e distingue tra modelli unidimensionali, che ipotizzano (o comunque prendono in considerazione) l'operare di una sola dimensione latente nella riproduzione dei *pattern* di risposta, e modelli multidimensionali, che indagano un numero maggiore di dimensioni, solitamente considerate tra loro indipendenti[27].

Un altro aspetto da cui partire per classificare i modelli di *scaling* è considerarne le fattispecie sulla base delle operazioni logicamente connesse alla rilevazione empirica della proprietà e quindi alla misura stessa. Bruschi [1999], per esempio, si occupa più diffusamente di cinque tipi principali di scale[28]. Due di queste sono interpretate come misurazioni fondamentali: scale di valutazione e cumulative. Le scale di *valutazione* sono basate sull'assegnazione di punteggi da parte di giudici a certi oggetti, in relazione al giudizio dell'intensità ordinale o numerica della proprietà posseduta. Si tratta, in sostanza, della stessa logica dell'assegnazione dei voti scolastici. Le scale *cumulative*, di cui si tratterà ampiamente più tardi parlando di quella di Guttman, ipotizzano proprietà con

[27] Un quinto criterio distingue tra modelli con un solo *item* e con più *item*. Nella pratica delle scienze psicologiche e sociali, la prima fattispecie è alquanto infrequente, per motivi di validità dello strumento di rilevazione: un solo *item* difficilmente è capace di discriminare la posizione di tutti i soggetti lungo un attributo.

[28] L'autore cita solamente, senza trattarle, le scale di comparazione, di preferenza (entrambe previste dalla classificazione di Coombs, che quindi è in parte complementare a quella di Bruschi), di discriminazione, della struttura latente.

strutture ordinate, per cui un soggetto coerente che accetta un *item* a un certo livello *x*, accetterà anche tutti gli altri *item* di livello più basso. Perciò, nel modello ideale sapere che una persona ha "passato" *almeno* il livello *x*, implica dedurre che abbia passato anche tutti quelli di minore intensità.

Altri tre tipi di scale sono invece, secondo l'autore, derivate, perché fondate su altre "misure" precedenti. I modelli di *scaling additivi* costruiscono la stima della posizione di un soggetto su un *continuum* attraverso la somma dei singoli punteggi ottenuti su una pluralità di *item*. Le scale Likert ne sono un esempio: il punteggio totale dell'individuo *X* è la somma dei singoli punteggi che esprimono il rispettivo grado di accordo con una batteria di affermazioni concernenti l'attributo in oggetto. Le scale a *intervalli apparentemente uguali* di Thurstone, di cui si parlerà in seguito, sono un tentativo di misurazione intervallare che si basa su un'elaborata procedura preliminare per collocare lungo una dimensione concettuale una serie di affermazioni in modo equidistante, a intervalli uguali secondo il parere di un *pool* di giudici. Le frasi con cui concordano, costituiscono la base della stima della posizione degli individui su quel *continuum*. Le scale *fattoriali*, infine, sono fondate su gruppi "d'indicatori fortemente correlati tra loro, quindi più garantiti nella loro unidimensionalità e validità teorica" [ibid, 110].

Una distinzione senz'altro cruciale ai fini della nostra ricostruzione è quella tra modelli di tipo deterministico e modelli di tipo probabilistico (o stocastico). Alle spalle della generalità dei modelli deterministici classici, vi è l'approccio tradizionale dei test psicometrici[29], per il quale l'obiettivo della stima è la determinazione puntuale del *punteggio vero* (*true score*) dello stato sulla proprietà che la variabile intende rilevare, minimizzando l'errore di misurazione. Il punteggio vero è un ideale, non osservato e non osservabile, ricorda Corbetta [2003, 52]. Ciò che rileviamo è il *punteggio osservato* (*observed score*), quello ricavato operativizzando il concetto iniziale, cioè stabilendo e mettendo in atto le operazioni pratiche di rilevazione. La relazione tra il valore osservato, quello vero e l'errore è data da:

[29] Una panoramica sulla teoria dei test va al di là di qualsiasi sforzo di sintesi. A titolo meramente indicativo citiamo Scott [1968], Frudà [1975], Nunnally [1978], Roberts [1979], Rubini [1984], Nunnally e Bernstein [1994], Pedon [1995], Barbaranelli e Natali [2005], Primi e Chiesi [2005], Boncori [2006].

$$O = T + e, \qquad \text{(eq. 2.1)}$$

dove *O* rappresenta il punteggio osservato, *T* il punteggio vero, *e* l'errore di misurazione. Se l'errore non è affetto da elementi di sistematicità (distorsioni) e può essere considerato casuale, allora la teoria assume che la sua distribuzione in *n* prove ripetute si avvicinerà a una curva normale con media pari a zero, al crescere di *n*. Questo vuol dire che al crescere del numero delle misurazioni, il punteggio osservato atteso coinciderà con quello vero.

I modelli stocastici non vanno alla ricerca *del* punteggio vero ma stimano piuttosto la probabilità di risposta associata a ogni valore del *continuum*, incorporando quindi nei propri assunti la dimensione aleatoria della variabilità delle risposte, soprattutto elaborando tecniche per la stima dell'adattamento del modello ai dati effettivamente osservati.

2.2 Commento

Tantissime sono le classificazioni dei modelli di *scaling*, delle tecniche, dei tipi di scala e dei criteri con cui poterli distinguere[30]. Il repertorio introduttivo presentato in questa sede ha un valore esemplificativo, teso a mettere in luce solo alcuni elementi rilevanti, quelli strettamente essenziali a inquadrare i passi successivi. Innanzitutto è fondamentale fissare l'attenzione sul fatto che la scelta di un modello e di una tecnica di *scaling* non è mai neutra, ma presuppone un *set* di assunti e ipotesi su ciò che si sta andando a *scalare*. Ogni criterio di distinzione tra modelli, ogni *fundamentum divisionis* dietro alle varie classificazioni presuppone in qualche modo un'idea sulla natura dei costrutti da "misurare". Da un certo punto di vista, richiamandoci al tema dell'ordine logico tra funzioni di assegnazione e strutture empiriche, la costruzione dello strumento costituisce e delinea il costrutto stesso. Il modello passa al vaglio della corroborazione empirica, pur essendo talvolta assai complesso distinguere, in caso di scarso adattamen-

[30] Altri esempi sono Coombs [1964], che sviluppa la classificazione di Stevens, cercando di aumentarne e specificarne le fattispecie: il caso delle scale metriche ordinali sarà discusso tra poco. Marradi [1981; 1985] propone una classificazione basata sulla natura delle procedure di rilevazione; Ricolfi [1985] risponde con una tipologia che tiene conto delle operazioni matematiche legittime per ogni tipo di scala.

to del modello ai dati, quale/i assunto/i sia/no inadeguato/i e quindi in cosa modificare il quadro. Senza contare che spesso mutare o eliminare o aggiungere uno o più assunti significa revisionare profondamente il concetto stesso in via di misurazione.

Ogni modello interpreta anche il rapporto tra la struttura latente che s'intende misurare e gli *item* che sono utilizzati come sua manifestazione. Si possono immaginare i secondi come espressione di una struttura ordinata e cumulativa fortemente integrata, oppure come elementi che insieme ricostruiscono i vari aspetti dell'intensione di un concetto[31], oppure immaginarli connessi da una rete di relazioni di preferenza ecc. Parimenti cambia anche la relazione che s'ipotizza tra la struttura latente (l'atteggiamento, il valore, la propensione ecc.) e i soggetti.

Infine, non dovremmo dimenticare, anche se la letteratura tecnica tende a metterlo tra parentesi, il fatto che ogni modello considera, spesso implicitamente, procedure e tecniche di rilevazione di un certo tipo in modo preferenziale, mentre è meno adatto a interpretare e processare materiale empirico operativizzato e codificato in modi diversi da quello considerato canonico. Ciò significa che ogni modello tende, quindi, ad abbinarsi ad alcune tecniche di rilevazione, che sono le più adatte a rappresentare l'oggetto cognitivo nella forma richiesta dagli assunti e dai caratteri generali del modello. Possono esserci casi in cui le informazioni rilevate sono non interpretabili o poco utili rispetto al taglio epistemologico previsto dal particolare modello; allo stesso tempo, da una medesima base empirica, dati rilevati con una specifica tecnica (per esempio di *rating*) possono essere mutati *ex post* in dati di tipo diverso (per esempio risposte dicotomiche). Questo è un aspetto potenzialmente critico, che chiama in causa il ruolo delle tecniche non solo, com'è ovvio, nel rilevare il dato, ma anche nell'influenzarlo, al di là dell'intervento del costrutto. L'impatto della tecnica può provocare fenomeni noti in letteratura sotto varie forme ("effetto alone", "*response set*", "fattore dello strumento" ecc). Campbell e Fiske [1959] focalizzano questo tema attraverso lo schema dei controlli multitratto-multitecnica, in cui oltre al tradizionale criterio di *convergenza* nella rilevazione dello stesso costrutto con tecniche diverse (i risultati devono correlare indipendentemente dalla procedura utilizzata), se ne prevede anche uno di *divergenza* nella rilevazione di costrutti diversi con la stessa tecnica (i risultati non devono correlare indi-

[31] E' questa una concezione molto diffusa del rapporto tra indicatori e concetto indicato [Marradi 2007].

pendentemente dal costrutto rilevato). La violazione di questo secondo criterio prefigura una situazione in cui la procedura orienta gli esiti delle risposte, piegando perciò i risultati sulla base delle caratteristiche della tecnica specifica: questo determina un eventuale circolo vizioso in cui l'immagine finale è proiettata, in misura maggiore o minore, dal modello e dalla tecnica stessi, più che dalla base empirica. Tale fattispecie, tra l'altro, è tutto fuorché infrequente o eccezionale[32].

[32] Un esempio (senz'altro tra tutti il più modesto) di studio empirico degli effetti sulle risposte dei soggetti della somministrazione parallela di *item* identici con tecniche diverse, è contenuto in La Sala [2010]: lì si mostrano vari tipi d'impatto, imputabili alla differente chiusura delle risposte, sul modo in cui i soggetti interpretano i quesiti e affrontano i processi di risposta.

3 Esempi classici di *scaling*: Thurstone, Coombs e Guttman

Dalle osservazioni svolte alla fine del capitolo precedente, che tracciano alcuni degli elementi essenziali da cui non è possibile prescindere quando si affronta un procedimento di *scaling*, faremo adesso un passo avanti in direzione del tema centrale della riflessione. Prima ancora di inoltrarci, nella Parte Seconda, nella trattazione del modello di Rasch, giova applicarsi brevemente su tre casi particolarmente indicativi di modelli di *scaling*. Tali esempi sono utili non solo perché mostrano, su un piano deterministico, elementi che ritroveremo trattando i modelli probabilistici della *Item Response Theory*, ma anche perché affrontano la questione dello s*caling* con approcci differenti: la ricerca della misurazione intervallare, di una metrica ordinata senza unità di misura e di una perfetta scalabilità cumulativa. Si tratta di un esercizio prezioso di studio, propedeutico al prosieguo del testo.

3.1 Thurstone: la legge del giudizio comparativo e le scale degli stimoli

Lo psicologo Louis Leon Thurstone fu uno dei pionieri della rilevazione sistematica di atteggiamenti e opinioni, tema cui si dedicò nel corso di un lungo arco temporale [Thurstone 1927, 1928, 1957, 1959; Thurstone e Chave 1929].

Nel suo articolo *Attitudes can be measured* [1928], pietra miliare delle nostre discipline, Thurstone offre alcune definizioni di atteggiamenti e opinioni che sono divenute classiche nel dibattito successivo delle scienze sociali. Scrive:

> il concetto di "atteggiamento" sarà qui utilizzato per denotare la somma totale delle inclinazioni e dei sentimenti di una persona, il pregiudizio o la preferenza, le nozioni preconcette, le idee, le paure, le minacce e le convinzioni a proposito di qualsiasi argomento specificato. (…) Il concetto di "opinione" significherà qui un'espressione verbale di un atteggiamento. (…) Il termine "opinione" sarà limitato

> all'espressione verbale. Ma è un'espressione di che cosa? Esso esprime, presumibilmente, un atteggiamento. (…) Un'opinione simboleggia un atteggiamento. Il punto successivo riguarda che cos'è che vogliamo misurare. (…) E' l'atteggiamento ciò che davvero ci interessa. L'opinione ha interesse solamente nella misura in cui la interpretiamo come un simbolo di un atteggiamento. Pertanto, ciò che vogliamo misurare è qualcosa di relativo agli atteggiamenti. Useremo le opinioni come mezzi per misurare gli atteggiamenti. [ibid., 531-2]

Alla base del contributo di Thurstone troviamo l'idea fondamentale che gli atteggiamenti siano interpretabili come *proprietà latenti* (non operativizzabili), di cui le espressioni verbali (ma anche comportamentali) sono manifestazioni[33] (indicatori operativizzabili). Tali opinioni, sulla base del grado dell'atteggiamento estrinsecato, possono essere collocate lungo un *continuum* concettuale che va da un minimo d'intensità dell'atteggiamento a un massimo.

Lo sforzo scientifico di Thurstone si indirizza, a questo punto, alla ricerca dei procedimenti più adeguati per collocare le opinioni lungo i possibili *continua* attitudinali[34]. Nelle scienze fisiche, l'attribuzione di un oggetto a una scala sulla base di una proprietà pone minori problemi teorici: ordinare una serie di gravi sulla base della loro massa, ad esempio, è facilmente conseguibile pesandoli uno a uno e registrando il loro valore. La situazione si complica notevolmente quando si ha a che fare con i giudizi soggettivi delle persone e con la mancanza di scale di misurazione di tipo fisico. Questo vale nel caso in cui volessimo far ordinare gli stessi gravi sulla base delle impressioni di pesantezza registrate da una serie di individui; ma vale a maggior ragione nel caso di *item* che rappresentano opinioni, non avendo in questo caso un raffronto "esterno" di tipo maggiormente "oggettivo". Una possibile soluzione è di proporre gli stimoli a coppie, chiedendo

[33] Thurstone è consapevole che le opinioni, come i comportamenti, possono essere indicatori non sempre validi degli atteggiamenti sottostanti, poiché ciò che si dice o si fa può non essere fedele a ciò che si pensa. Ma questo è un problema universale, sostiene, che appartiene a qualsiasi tentativo di rilevare le dimensioni umane e fisiche. Si veda anche Diana [2008].

[34] Giampaglia rileva che "l'obiettivo che egli perseguì instancabilmente per tutta la vita – avanzando varie proposte in tempi diversi, spesso modificate per migliorarle, accogliendo a volte anche i correttivi suggeriti da altri studiosi – consisteva nel tentativo di ordinare su un continuo psicologico degli oggetti (o stimoli o *item*), tenendo conto della quantità di proprietà (o attributo) posseduta da ciascuno di essi." [2008, 9].

di volta in volta quale sia il più pesante e ricostruendo così in via indiretta un ordinamento. Quest'ultimo tipo di schema generale è vicino a ciò che succede nel campo della misurazione psicologica [McIver e Carmines 1981, 17].

3.1.1 La legge del giudizio per confronto

Thurstone elabora un principio generale che chiama "legge del giudizio per confronto". Si ipotizza che in *ogni individuo*, posto di fronte a un *certo stimolo* da giudicare in conformità a un *criterio definito* e in un *dato momento*, operi un meccanismo chiamato *processo discriminale*, che "rappresenta l'*impressione* che riceve (…), la sua reazione quando gli viene chiesto di esprimere un giudizio" [Giampaglia 2008, 10]. Il tipo di processo discriminale attivato può dipendere senz'altro, in primo luogo, da un'ipotetica e astratta posizione *vera* dell'*item* lungo l'arco delle possibili intensità di un atteggiamento. In gioco entrano però anche tutti gli altri elementi appena evidenziati che, presi individualmente e tenendo ipoteticamente fermi di volta in volta tutti gli altri fattori, possono attivare processi discriminali distinti e diversi tra loro. Individuo, *item*, proprietà coinvolta, tempo sono variabili che possono condurre a una certa mutabilità nel giudizio sulla posizione dello stimolo lungo il continuo psicologico studiato. Per questo, registrando i giudizi sullo stesso *item* dati nel tempo da un medesimo soggetto o nello stesso momento da una pluralità di soggetti, si osserverà prevedibilmente una certa dispersione dei punteggi. A questo proposito "la legge dei giudizi comparati presume che per ogni stimolo (…) esista una risposta che occorre più frequentemente, alla quale ci si riferisce come *processo discriminale modale*" [McIver e Carmines 1981, 17]. Tale valore modale diventa quello dell'*item* lungo la dimensione dell'atteggiamento.

Thurstone assume che i processi discriminali per ogni stimolo si distribuiscano secondo una curva normale, caratterizzata da un certo valore centrale (su cui convergono media, moda e mediana) e da una certa varianza, che rappresenta la *dispersione discriminale* dello stimolo. Ma i valori dei processi discriminali non sono noti sin dall'inizio. Sfruttando le caratteristiche delle curve normali, si giunge a poter stimare i *giudizi discriminali*, cioè le differenze tra i processi discriminali relative a ogni coppia di stimoli, sotto forma di punteggi *z*, ossia le unità standardizzate della distribuzione normale. In particolare, il punto *z* di interesse è quello che si trova lungo l'ascissa, in corrispondenza del limite dell'area sottostante la curva normale, pari alla proporzione di

casi che hanno giudicato uno dei due *item* maggiore dell'altro rispetto alla proprietà di interesse. Il principio teorico che giustifica tale procedimento sulle differenze è dato dall'assunto che "il grado nel quale qualsiasi due stimoli possono essere discriminati è una funzione diretta della differenza nella loro posizione considerata riguardo all'attributo in questione" [ibid., 18].

Per giungere al risultato che la differenza tra il processo discriminale modale dello stimolo A e quello dello stimolo B è uguale al valore di *z* nel punto in cui la superficie sotto la curva normale è uguale alla proporzione di individui che hanno ritenuto A maggiore di B, servono altri assunti in grado di semplificare il quadro. In particolare, per raggiungere questo caso specifico Thurstone assume non solo che i processi discriminali si distribuiscano come una curva normale (e quindi anche le loro differenze si distribuiscano analogamente), ma anche che tutti i processi discriminali abbiano uguale varianza e siano tra loro completamente indipendenti, cioè il loro coefficiente di correlazione sia nullo. Questi assunti rendono particolarmente stringenti le condizioni teoriche cui è necessario sottostare affinché il modello possa essere considerato valido, costituendo un limite comune a tutte le tecniche sviluppate da Thurstone sulla base della legge del giudizio per confronto.

Tra queste, che restano alcuni dei tentativi più rigorosi e interessanti di costruire scale a intervalli, le più note in letteratura sono tre: la tecnica dei confronti a coppie (*paired comparisons*), la tecnica degli intervalli apparentemente uguali (*equal-appearing intervals*) e la tecnica degli intervalli successivi (*successive intervals*). In tutte queste tecniche i giudici (i soggetti preposti alla valutazione degli *item*) devono esprimere un giudizio *non* riferito al loro grado personale di accordo o disaccordo verso le affermazioni, bensì solamente alla posizione dell'*item* lungo il *continuum* dell'atteggiamento verso l'oggetto in esame: in uno studio su uomini politici, per esempio, in cui si devono collocare i personaggi lungo una dimensione "liberalismo/conservatorismo", non si chiede di considerare quanto ognuno di loro sia più o meno liberale o conservatore rispetto al giudice, ma dove quest'ultimo pensa che si collochi ogni politico su quella dimensione. Alcuni autori [Scott 1968] hanno mostrato quanto quest'assunto possa essere facilmente violato, in virtù del fatto che i giudici non possono spogliarsi della posizione che loro stessi ricoprono sulla dimensione in oggetto: in

particolare, individui con posizioni estreme sul *continuum* tendono ad avere minor capacità di discriminare correttamente *item* collocati in posizioni più centrali.

3.1.2 La tecnica dei confronti a coppie

La prima tecnica (*paired comparisons*) si basa sul confronto di tutte le possibili coppie di *item*. Per ciascuna di queste si chiede a ogni soggetto di selezionare, tra i due, lo stimolo con la quantità di proprietà maggiore. Si prosegue poi applicando la procedura descritta in precedenza per il calcolo dei punteggi, sulla base delle proporzioni di risposta riportate sulla curva normale. In questo modo, assumendo rispettate tutte le caratteristiche teoriche del modello, si riproducono le distanze tra ogni coppia, mantenendo invariata la capacità di discriminare in modo uguale ciascuna di esse. Vi è però un limite pratico molto consistente: l'insieme delle possibili combinazioni di tutti gli *n item* è pari a $n * (n - 1) / 2$. Giacché è necessario [Edwards 1957] utilizzare un vasto repertorio di frasi per coprire tutto l'arco delle possibili posizioni lungo il continuo psicologico che stiamo studiando, gli individui incaricati di valutare le coppie si trovano di fronte a un compito arduo: se limitiamo la batteria a 15 affermazioni, per esempio, essi devono occuparsi di giudicare $15 * 14 / 2 = 105$ coppie, che salgono a 190 nel caso di 20 *item* e 300 nel caso di 25. Si capisce come una tale incombenza possa, nel corso di un test, comportare dei cali notevoli di concentrazione e di capacità discriminante.

3.1.3 La tecnica degli intervalli apparentemente uguali

Nella tecnica degli *intervalli apparentemente uguali*, ai soggetti è consegnato un *set* solitamente consistente di affermazioni, con il compito di distribuirle in un certo numero di pile (solitamente dispari) sulla base del grado di favore che esse esprimono verso l'oggetto in questione: la prima pila rappresenta il massimo di favore, l'ultima il massimo di sfavore, quella centrale una posizione neutra. L'elemento fondamentale di tale tecnica è che i giudici devono distribuire gli *item* in modo tale che le distanze tra le categorie siano percepite come uguali tra loro. Si tratta del tentativo per eccellenza di costruire scale che, affiancando intervalli uguali, riprendano la struttura di quelle metriche e si collochino al livello di misurazione intervallare. Del resto, come annunciato anche dal nome, tale uguaglianza è solo apparente, in quanto approssimata [Bruschi 1999,

108] e affidata al processo percettivo e discriminante del giudizio soggettivo. Infatti, il modello non fornisce strumenti per sottoporre a controllo l'assunto dell'uguaglianza degli intervalli. Al contrario, è possibile mostrare [Edwards 1957, 120-3] che agli estremi del continuo con molta probabilità si assisterà a uno schiacciamento delle categorie, poiché *item* via via più estremi andranno a finire sempre nella prima e nell'ultima pila, senza che necessariamente le affermazioni relativamente meno estreme siano continuamente ridistribuite nelle categorie più centrali, ricalibrando l'insieme degli *item* già sistemati[35]. In generale è comunque arduo poter sostenere il parallelismo con la forma metrica di rilevazione, poiché ciò significherebbe assumere che nella mente di ogni giudice gli intervalli possiedono effettivamente la stessa distanza: ma nessun confronto intersoggettivo di tale sorta è empiricamente possibile. A prescindere da tali considerazioni, una volta che ogni giudice ha disposto le affermazioni nelle pile, si prosegue calcolando per ognuna di esse la categoria mediana di assegnazione e la varianza: esse assumono rispettivamente il significato di processo discriminale modale e di dispersione discriminale. Per costruire la scala finale si cerca di ridurre il numero degli *item*, seguendo alcuni criteri. Per esempio, si selezioneranno in maniera da rappresentare tutte le posizioni del *continuum* in modo il più possibile perequato e si elimineranno le affermazioni con un'eccessiva dispersione discriminale.

3.1.4 La tecnica degli intervalli successivi

La terza tecnica è quella degli *intervalli successivi*. Inizialmente si procede come per gli intervalli apparentemente uguali, chiedendo di distribuire le affermazioni nelle categorie precostituite. Stavolta, però, l'unico vincolo posto è quello del rango: le pile rappresentano intervalli ordinati che però non devono essere necessariamente considerati equidistanti. Per ogni affermazione si prendono in considerazione le rispettive distribuzioni nelle varie categorie e si calcolano le proporzioni cumulate di risposta per ogni soglia. Assumendo che le risposte si distribuiscano come una curva normale, sono cal-

[35] Se incontro un'affermazione che giudico la più sfavorevole possibile all'oggetto, la porrò nell'ultima pila; se in seguito ne incontro una che ritengo ancora più estrema, essa finirà di nuovo nell'ultima pila e sarà considerata uguale a quella precedente (da qui lo schiacciamento), salvo che non decida di operare, in considerazione del nuovo limite estremo, una nuova risistemazione complessiva di tutti gli *item* già collocati. Il che, considerando che questi potrebbero essere svariate decine, complica notevolmente il compito.

colati i rispettivi valori di *z* per ogni categoria. La differenza di questi ultimi tra categorie adiacenti determina l'ampiezza degli intervalli per ogni *item*. Di solito si utilizza la media dell'ennesimo intervallo su tutti gli *item* per stabilirne un'ampiezza comune. Rispetto alla tecnica precedente, quella a intervalli successivi non soffre dello schiacciamento degli intervalli estremi (il c.d. *end-effect* [Giampaglia 2008, 21]) e la stima *a posteriori* dell'ampiezza appare una forma di quantificazione più verosimile rispetto all'assunto *a priori* che essi siano uguali per tutti. Dall'altro lato, è impossibile stimare i confini per le due categorie estreme, che non hanno limiti nella parte inferiore e in quella superiore della scala. Permangono chiaramente altre criticità, già sottolineate, comuni a tutte le tecniche scaturite dal modello della legge dei giudizi per confronto: in particolare l'assunto della separabilità tra giudizio espresso dai giudici e loro posizione e l'assunto della normalità della distribuzione.

Il contributo di Thurstone è stato il primo tentativo *forte* di costruire un modello organico per la misurazione degli atteggiamenti, fatto di assunti teorici, ipotesi sul funzionamento del processo cognitivo di espressione dei giudizi ed elaborazione di tecniche orientate a fornire i dati necessari. L'orientamento del modello è chiaramente rivolto agli *item*, ossia alla costruzione dello strumento che serve a rilevare gli stati di un soggetto su una proprietà. Uno dei limiti è che la calibratura dello strumento richiede il giudizio dei soggetti stessi, lasciando aperto il rischio che la distribuzione della proprietà in questione tra i giudici influenzi la struttura dello strumento. In generale, i modelli di *scaling* proposti non sembrano giustificare in modo completo un'interpretazione di tipo metrico e intervallare, pur rappresentando un passo necessario nel chiarire cosa significhi tale livello di misurazione nelle scienze sociali.

3.2 Il modello *unfolding* e le metriche ordinate

Il modello elaborato dagli anni '50 da Coombs si pone su un livello di "misurazione" a cavallo tra quello ordinale e quello a intervalli[36]. L'originalità di questo contri-

[36] La ricostruzione della teoria di base, esposta in questa sezione si basa essenzialmente su Coombs [1950, 1964] e Mc Iver e Carmines [1981].

buto ci spinge a dedicare le prossime pagine all'illustrazione degli elementi teorici della versione più semplice del modello, quella unidimensionale. La logica generale di questo approccio è molto semplice: dall'ordinamento di una serie di oggetti lungo una dimensione, è possibile indagare se dietro le singole scelte individuali si possa rintracciare una scala generale comune con cui le prime siano compatibili. Inoltre, pur non facendo "nessuna assunzione sull'uguaglianza degli intervalli [il modello mostra che] sulla base di assunti accettabili e con la tecnica appropriata siamo in grado di ordinare la grandezza degli intervalli tra gli oggetti" [Coombs 1950, 145].

Coombs apre l'articolo che introduce il suo modello definendo la misurazione come "l'assegnazione di numeri a oggetti con la condizione che questi numeri obbediscano alle regole dell'aritmetica" [ibid.]. Anche se era abitudine comune ritenere che le scale usate in psicologia fossero a intervalli, con un'unità di misura fissa, Coombs riconosce che

> questo tipo di scala dovrebbe essere usata solo se può essere sperimentalmente dimostrato attraverso la manipolazione degli oggetti che i numeri assegnati agli oggetti obbediscono alle leggi dell'addizione. L'unità di misura in psicologia, tuttavia, è ottenuta da una combinazione di definizioni e assunti i quali, se visti come una prima approssimazione e associati con una teoria statistica dell'errore, servono molti propositi pratici. Ma poiché possiamo talvolta mettere in discussione il significato delle definizioni e la validità degli assunti che conducono a un'unità di misura, è nostra intenzione in questo articolo sviluppare un nuovo tipo di scala che non comporti un'unità di misura. [ibid.]

3.2.1 Scale individuali e scala congiunta

Il modello utilizza i dati che lo stesso Coombs definisce di *scelta preferenziale* che, abbiamo visto, sono ottenuti attraverso l'ordinamento da parte di un soggetto, secondo un criterio, di una serie di oggetti o *item*: solitamente il criterio è quello della preferenza individuale, per cui si classificano gli oggetti da quello preferito a quello meno.

Ogni soggetto intervistato elenca gli *n item* in un dato ordine[37], chiamato *scala I* (dall'inglese *Individual Scale*, scala individuale). Il punto è che un insieme di scale I, seppur diverse tra loro a prima vista, può essere complessivamente compatibile con quella che è chiamata *scala J* (dall'inglese *Joint Scale*, scala congiunta). Nel caso si verifichi questa fattispecie, è possibile concludere che esiste un attributo latente comune a ogni soggetto, sulla base del quale sono state operate le scelte di preferenza e che tale attributo è rappresentato proprio dalla scala congiunta. Al contrario si conclude che gli attributi latenti necessari a spiegare le scelte sono più di uno (passando così al modello multidimensionale) oppure che le scelte non sono influenzate da nessuna dimensione latente specifica.

Vi sono diverse soluzioni grafiche per mostrare la logica soggiacente al confronto tra le scale I e la scala J[38]. L'obiettivo è rappresentare sia gli stimoli che gli individui "in uno spazio unidimensionale tale che le distanze relative tra loro riflettano la prossimità psicologica degli stimoli agli individui" [McIver e Carmines 1981, 72]. Una rappresentazione lineare molto semplice aiuta a capire la base del ragionamento e in che modo scale individuali diverse possano ricomporsi in una sola scala congiunta. Si osservi la figura seguente:

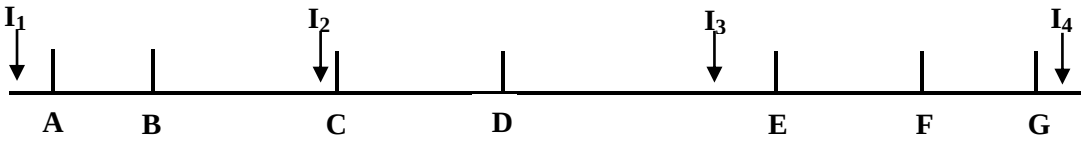


Fig. 3.1 – Rappresentazione di una dimensione latente con 7 *item* e 4 soggetti

S'immagini che la linea retta orizzontale rappresenti la dimensione ipotetica su cui si trovano 7 stimoli/*item* (da A a G) e 4 individui (da I_1 a I_4), posti come indicato rispettivamente dalle tacche e dalle frecce. La scala J è data dall'ordine ABCDEFG degli *item*. Da tale ordinamento e dalla posizione dei quattro soggetti, scaturiscono altrettante scale di preferenza individuali compatibili con quella congiunta. Per l'individuo I_1, posto all'estremo inferiore della scala, l'ordinamento è esattamente ABCDEFG;

[37] Un insieme di *n* elementi può essere combinato in *n*! modi diversi, una prima metà dei quali specolari rispetto alla seconda metà.

[38] Prende il nome di *unfolding* (dispiegare, svelare, aprire, allargare) proprio la procedura con la quale si confronta la coerenza delle scale I con la scala J.

l'individuo I_4, collocato all'estremo opposto, "vede" gli stimoli in modo speculare: GFEDCBA; l'ordinamento compatibile dell'individuo I_2 è CBDAEFG; l'individuo I_3, infine, ha ordinamento EFDGCBA. Il processo descritto, che a partire dalla scala J passa a determinare l'ordine degli stimoli per gli individui sulla base della loro posizione sulla dimensione, è chiamato *folding* della scala J sulle scale individuali. Il processo inverso, quello di *unfolding* degli ordinamenti individuali per formare una scala J, parte dall'osservazione delle quattro scale I (ABCDEFG, CBDAEFG, EFDGCBA, GFEDCBA) e ricerca, se esiste, l'ordinamento con loro compatibile (in questo caso ABCDEFG).

3.2.2 Interpretazione qualitativa o quantitativa della scala congiunta

La teoria mostra che esistono due tipi di scale J per ogni *set* di scale I. Infatti, possiamo essere interessati solo al livello ordinale della scala ovvero possiamo rivolgere l'attenzione a sviluppare anche le sue caratteristiche di *metrica ordinata*. Nel primo caso si riflette su un certo numero di scale I e si cerca una scala J, compatibile con esse, che rappresenti solo gli stimoli come una serie ordinata da un'estremità all'altra: in questo caso si definisce la scala J di tipo *qualitativo*. Non ci si pone la questione dell'ordine di grandezza delle distanze, ma solo quella del loro ordine di rango. Si può dimostrare che a ogni scala qualitativa congiunta di *n* stimoli corrispondono 2^{n-1} scale individuali compatibili con essa.

Nel caso, invece, in cui si vogliano considerare, oltre all'ordine degli *item*, anche le informazioni riguardanti la loro distanza, si darà della scala J un'interpretazione *quantitativa*. Ciò significa che, attraverso la tecnica opportuna, si potrà controllare se le scale individuali empiricamente rilevate formano o no un insieme compatibile con l'ipotesi che la distanza tra alcuni degli stimoli sia maggiore o minore della distanza tra altri stimoli. I criteri sono ovviamente più stringenti rispetto al caso della scala J qualitativa. Non rientra nei nostri obiettivi un'illustrazione completa e puntuale della tecnica, per cui ci limiteremo qui a esporre sommariamente i punti principali.

Una scala congiunta quantitativa è composta di un sottoinsieme delle 2^{n-1} scale I compatibili con la corrispondente scala J qualitativa: il numero massimo di scale indivi-

duali di cui è composto tale sottoinsieme è $\binom{n}{2} + 1$ [39]. Per questo, una scala J con 4 stimoli sarà compatibile con 8 scale I nella versione ordinale e con 7 scale I nella versione metrica ordinata; una scala J con 5 stimoli sarà compatibile con 16 scale I nella versione ordinale e con 11 scale I nella versione metrica ordinata, e via dicendo. Il numero di stimoli deve essere superiore a tre per ricavare informazioni sull'ordine di grandezza delle distanze. Dato un insieme di *n item*, è possibile ricavare una singola informazione sull'ordinamento di due distanze per ogni sottoinsieme disponibile di 4 *item*: il numero di sottoinsiemi di 4 elementi ricavabili da un insieme di *n* è dato dal calcolo combinatorio come $\binom{n}{4}$ [Torgerson 1958].

Per passare dall'ordine semplice all'ordine tra differenze, il modello ha bisogno di un assunto in realtà abbastanza "forte", che introduce la dimensione metrica: si suppone, infatti, che la distanza tra due *item* sia in funzione diretta dell'intensità della loro preferenza e che esista un punto centrale tra i due stimoli che rappresenta una posizione di equidistanza. Rilevare empiricamente certe scale I e non altre può fornire informazioni sulla posizione relativa di taluni punti centrali e, a sua volta, sulla distanza relativa tra gli stimoli coinvolti.

Vedremo ora un esempio riferito alla fattispecie più semplice (4 stimoli)[40].

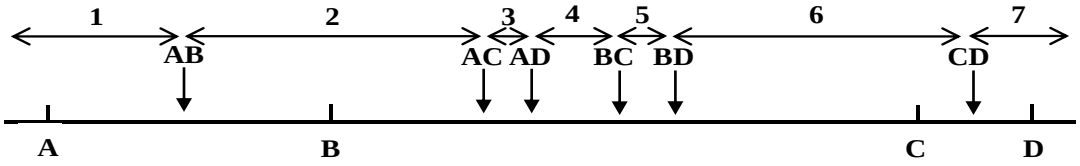


Fig. 3.2 – Rappresentazione di una dimensione latente con 4 *item* e $\overline{AB} > \overline{CD}$

[39] Vi sono altre tre caratteristiche *necessarie* affinché una scala J sia quantitativa [McIver e Carmines 1981, 75-6]. Per le nostre esigenze sarà sufficiente citarle in nota:

- Ogni scala I deve necessariamente terminare o con il primo o con l'ultimo stimolo della scala J.
- Può esserci solo una coppia di scale I le quali siano una l'immagine esattamente speculare dell'altra. Una di esse comincerà con il primo stimolo della scala J e finirà con l'ultimo; viceversa per quella a essa speculare.
- Se un insieme completo di scale I (composto cioè da $\binom{n}{2} + 1$ scale) si combina con la scala J, deve essere possibile sistemare le scale individuali in modo tale che ognuna di esse scaturisca da quella precedente, con la sola inversione di una coppia adiacente di *item*.

[40] L'esempio è tratto da McIver e Carmines [1981]. Esempi più complessi si possono trovare nelle opere di Coombs citate. In particolare, Coombs [1950] fa alcuni esempi con scale di 5 stimoli.

S'immagini che i quattro *item* siano disposti lungo l'attributo rappresentato dall'asse orizzontale, nell'ordine ABCD. Le frecce verso il basso indicano i sei punti centrali tra gli stimoli, che separano il continuo in sette intervalli caratterizzati ognuno da un certo ordine di preferenze (frecce orizzontali in alto), ricostruibile come segue:

1) ABCD
2) BACD
3) BCAD
4) BCDA
5) CBDA
6) CDBA
7) DCBA

Come si vede, ogni volta che si supera una soglia di equidistanza tra due *item,* si passa a un ordinamento diverso delle preferenze e, superando, si inverte l'ordine di questi ultimi.

Quante e quali informazioni metriche possiamo ottenere dai quattro stimoli? Come detto, esse sono pari a $\binom{n}{4}$, in questo caso 1. Nell'esempio precedente, l'informazione che si ricava è che l'intervallo AB è maggiore di quello CD. La scala decisiva è la quarta, BCDA. Essa ci dice che, partendo dall'estremità sinistra della dimensione, il punto centrale tra A e D si incontra prima di quello tra B e C. Infatti, la scala della terza sezione è BCAD, mentre in quella successiva si ha appunto l'inversione tra A e D. Il fatto che il punto di equidistanza di AD sia più a sinistra di quello di BC è garanzia che C sia più vicino a D di quanto B non lo sia ad A e quindi che $\overline{\mathrm{AB}} > \overline{\mathrm{CD}}$.

La possibile alternativa sarebbe che la quarta scala invertisse, rispetto alla terza, gli item B e C, presentandosi quindi come CBAD. In questo caso avremmo concluso che il punto centrale di AD si trova a destra di quello di BC e quindi che $\overline{\mathrm{AB}} < \overline{\mathrm{CD}}$. Graficamente la configurazione cambierebbe nel modo seguente:

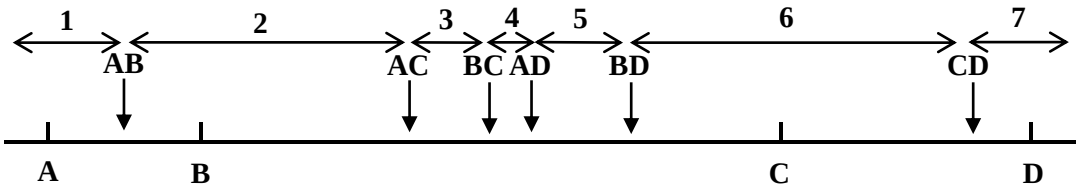


Fig. 3.3 – Rappresentazione di una dimensione latente con 4 *item* e $\overline{\mathrm{AB}} < \overline{\mathrm{CD}}$.

L'ordine degli altri *mid-point* non può cambiare, data la combinazione ABCD, per motivi strutturali: quello tra A e B deve essere logicamente prima di quelli tra A e C e tra A e D; il punto tra A e C deve essere prima di quello tra B e C, il quale a sua volta deve precedere quello tra B e D, che deve venire prima di quello tra C e D. Il "grado di libertà", se così vogliamo chiamarlo, riguarda quindi l'ordine tra il punto centrale di BC e quello di AD.

Non è facile, da un punto di vista pratico della ricerca applicata, che durante la rilevazione in un campione si abbia evidenza di sole scale individuali compatibili univocamente con una sola scala J. Questo è uno dei limiti comune a molti degli approcci deterministici: basta un solo caso incoerente per contraddire il modello e l'insieme particolare di assunti che lo costituiscono[41]. Inoltre, superando i 5 *item*, il modello acquisisce una complessità difficilmente gestibile e forse questo è uno dei motivi che ne hanno frenato la diffusione.

Il modello della metrica ordinata ha comunque molti elementi interessanti. Uno di questi è che si presta sia a un tipo di ricerca esplorativa che a una di tipo confermativo. Infatti, si può iniziare dai dati empirici per testare se essi siano coerenti con una qualsiasi rappresentazione dimensionale delle preferenze, senza un piano preciso, facendo in modo che sia la tecnica a fornire un'eventuale scala congiunta da interpretare e decifrare *ex post*; oppure si può partire da un'ipotesi teorica precisa, immaginando che esista una

[41] A tutti i modelli deterministici sono stati associati coefficienti e indici per stabilire la quota di "errore", di casi devianti dalla norma richiesta dal modello stesso. Tutte queste misure soffrono però di arbitrarietà nello stabilire quale quota di devianza sia da ritenere tollerabile e oltre la quale devono invece considerarsi violate a tal punto le caratteristiche del modello, da classificarlo come falsificato dai dati.

dimensione di preferenze strutturate come una certa scala J, controllando in seguito se le scale I sono compatibili con l'ipotesi iniziale o se essa sia falsificata dai dati.

E' inoltre notevole come il modello (se corroborato dalle evidenze empiriche) partendo da semplici dati di preferenza, riesca ad aggiungere una quantità d'informazioni non indifferente. Usando le parole di Coombs [1950, 157-8], esse possono essere

> 1) la scoperta di un attributo latente soggiacente alle preferenze; 2) l'ordine degli stimoli sul *continuum* dell'attributo; 3) qualcosa al riguardo delle grandezze relative delle distanze tra coppie di stimoli; 4) il frazionamento del *continuum* in intervalli, il posizionamento delle persone in questi intervalli e l'ordinamento di questi intervalli su questo *continuum* dell'attributo; 5) qualcosa a proposito delle grandezze relative di questi intervalli.

Per ottenere almeno parte di questi risultati, si ricorre però a un assunto, quello della relazione lineare tra intensità della preferenza e distanza degli *item*, che abbiamo definito "forte". Perché? Si sviluppa un approccio che fa dichiaratamente a meno di unità di misura perché, si dice, esse sono spesso definite in modo controverso e insoddisfacente nelle scienze psicologiche e sociali, che non possono manipolare i propri oggetti di ricerca e verificare così gli assunti di misurazione; si ricorre quindi a dati di tipo ordinale, quali appunto le preferenze, più verosimili; si continua, però, a considerare la dimensione dell'attributo latente come continua e quantitativa, non solo sullo sfondo, a livello di aspetto teorico generale e di contesto, ma come vera e propria indicazione operativa, poiché si utilizzano in modo diretto le caratteristiche che scaturiscono da questo assunto metrico: gli *item* hanno determinate distanze tra loro, incognite ma direttamente in relazione con l'intensità (quantitativa) della preferenza e tale caratteristica è considerata così precisa e strategica da poter immaginare che ognuna di loro sia perfettamente sezionabile a metà in un punto di equidistanza tra due stimoli, che determina a sua volta i confini degli intervalli che individuano le zone delle varie scale I.

Il modello *unfolding* di Coombs, nella versione base che abbiamo commentato, si colloca quindi a metà strada tra un livello ordinale e uno intervallare. A differenza della generalità dei modelli per scale ordinali, qui non si sospende il giudizio sulla natura

dell'attributo: essa è chiaramente interpretata come continua e quantitativa e tale caratteristica entra direttamente nella fattispecie delle scale J quantitative. Senza tale assunto, dalla semplice combinazione di relazioni di rango, non potrebbero emergere le altre, interessanti informazioni sull'ordine delle grandezze. Ma, al pari della definizione di una qualsiasi unità di misura per dimensioni psicologiche, anche quella delle relazioni intercorrenti tra stimoli lungo l'attributo, offerta dal modello, è ipotetica e ugualmente questionabile dai sostenitori di una visione classica della misurazione.

3.3 Il modello cumulativo di Guttman

A metà degli anni '40 Leon A. Guttman propone uno strumento che affronti in modo più efficace di quanto fatto sino a quel momento la questione dell'accertamento dell'*unidimensionalità* espressa da una serie di *item*[42]. Possiamo vedere il modello di Guttman come un precursore in chiave *deterministica* del modello del Rasch e ciò, in questa sede, rappresenta un altro motivo di attenzione.

Nel modello, gli *item* rappresentano vari gradi d'intensità di un atteggiamento verso un certo oggetto cognitivo. Le dimensioni attitudinali sono interpretate in modo *ordinale* e *cumulativo*: gli *item* sono unidimensionalmente scalabili se i) possono essere ordinati da quello che esprime un minimo di proprietà (più "facile", nel senso che quasi tutti sono disposti ad accettarlo) a quello che ne esprime un massimo (più "difficile", cioè pochi o al limite nessuno concordano con esso) e ii) i soggetti che concordano con un *item* posto a un certo livello, concordano anche con tutti gli *item* posti a un livello di "difficoltà" inferiore. Per valutare questi due aspetti, il modello considera congiuntamente gli stimoli e gli individui, ponendoli su una stessa dimensione. Idealmente, un soggetto che si colloca in un punto di una dimensione concettuale esprimerà un atteggiamento verso un dato oggetto con una certa forza: per esempio, il livello di tolleranza verso i membri di una certa etnia. La condizione di scalabilità richiede che egli accetti tutti gli *item* posti al suo stesso livello e a quelli inferiori sulla corrispondente scala, mentre rifiuterà quelli che richiedono una maggiore intensità nell'atteggiamento: per

[42] La letteratura sull'argomento è sterminata. Per la stesura di questa sezione si fa riferimento principalmente a Guttman [1944, 1950], McIver e Carmines [1981], Giampaglia [1990, 2008], Bruschi [1999], Marradi [2007].

proseguire l'esempio, accetterà gli *item* che corrispondono al suo livello di tolleranza e a quelli inferiori, mentre non sarà disposto a concordare con quelli che esprimono un livello di tolleranza superiore al suo[43]. La natura del modello e dei suoi assunti fa in modo che, in caso di perfetta scalabilità, sia possibile prevedere, sulla base dell'abilità di un soggetto, quali *item* egli accetti e quali invece rifiuti. Guttman definisce le scale partendo da questi elementi:

> Per una data popolazione di oggetti, la distribuzione di frequenza multivariata di un universo di attributi sarà chiamata una s*cala* se è possibile derivare dalla distribuzione una variabile quantitativa con la quale caratterizzare gli oggetti in modo tale che ogni attributo sia una semplice funzione di quella variabile quantitativa. Una variabile quantitativa siffatta è chiamata una variabile di scala. [Guttman 1950, 64]

3.3.1 Lo scalogramma

Nel modello base[44] si chiede ai soggetti di valutare le affermazioni in modo dicotomico: "in accordo/in disaccordo", indicando l'accordo con il codice '1' (successo nel "superare" l'*item*, il soggetto raggiunge quella soglia) e il disaccordo con il codice '0' (insuccesso nel "superare" l'*item*, il soggetto non raggiunge quella soglia). Lo *scalogramma* è lo strumento focale per l'analisi nel modello di Guttman. Esso è una tabella a doppia entrata che rappresenta nello stesso spazio i soggetti e gli *item*, con le risposte dei primi ai secondi. Lo scalogramma aiuta a comprendere e analizzare il grado di aderenza dei dati alla condizione di perfetta scalabilità. In quest'ultimo caso si presenterà come nella tabella seguente:

[43] Un precursore del modello di Guttman per l'utilizzo di questa logica è la scala di distanza sociale introdotta da Bogardus [1925], che utilizza affermazioni che descrivono comportamenti, considerati in relazione gerarchica tra loro rispetto a una dimensione teorica. Tanto maggiore è l'apertura verso un oggetto cognitivo contenuta in un comportamento, tanto minore è la distanza sociale da esso.

[44] Esistono anche versioni per *item* politomici.

	Item						
soggetti	**I_1**	**I_2**	**I_3**	**I_4**	**I_5**	**I_6**	punteggi
A	1	1	1	1	1	1	6
B	1	1	1	1	1	0	5
C	1	1	1	1	1	0	5
D	1	1	1	1	0	0	4
E	1	1	1	0	0	0	3
F	1	1	1	0	0	0	3
G	1	1	0	0	0	0	2
H	1	0	0	0	0	0	1
L	0	0	0	0	0	0	0
totali item	8	7	6	4	3	1	

Tab. 3.1 – Esempio di scalogramma di Guttman nel caso di perfetta scalabilità, con 9 *item* e 6 soggetti.

In riga sono riprodotti i profili di risposta di nove ipotetici individui, mentre in colonna le risposte a sei ipotetici *item*. I marginali di riga rappresentano la somma delle affermazioni accettate dai soggetti, il loro *punteggio* sulla scala; i marginali di colonna sono il totale dei soggetti che hanno accettato ciascun *item*. Il nostro esempio è un caso di *perfetta scalabilità* perché è possibile ordinare i soggetti e le affermazioni in modo tale che i marginali di riga e di colonna siano decrescenti: gli *item* sono ordinati da quello ritenuto complessivamente più accettabile dall'insieme del campione (8 soggetti su 9 lo hanno approvato) a quello meno accettabile (un solo soggetto lo condivide), mentre i rispondenti sono ordinati per punteggio, da quello che ha approvato tutti gli *item* (anche quello più "difficile") a quello che li ha rifiutati tutti (anche il più "facile"). Inoltre, ogni soggetto non ha alcun codice '0' alla sinistra di un '1', né alcun *item* ha un codice '0' sopra un '1'. In questo esempio abbiamo inserito due punteggi (5 e 3) cui corrispondono ognuno due soggetti: questo per indicare che all'interno dello scalogramma possiamo rilevare più persone (o *item*) con un punteggio uguale, i cui profili sono assolutamente indifferenti tra loro. Nel caso in cui avessimo avuto uno e un solo soggetto per ogni punteggio (da quello massimo a 0), lo scalogramma sarebbe diventato una matrice perfettamente triangolare, con tutte cifre '1' sopra la diagonale secondaria e tutte cifre '0' al di sotto.

Il concetto di scalabilità fa il paio con quello di *riproducibilità*: conoscendo il punteggio di un soggetto si è in grado di stabilire quanti e quali *item* abbia accettato. Più precisamente, se un individuo ha un punteggio pari a *n*, in caso di perfetta scalabilità si è in grado di affermare che egli ha accettato esattamente i *primi n item* (partendo da sinistra dello scalogramma, cioè dall'affermazione più "facile") e ha rifiutato tutti gli altri. Per questo si suole dire che, dalla conoscenza del punteggio, *l'intero profilo del soggetto è perfettamente riproducibile*.

Il punteggio assegnato a ogni soggetto sulla base delle risposte non ha alcun carattere di cardinalità, ma è solo un indicatore ordinale. Il modello di Guttman, in sostanza, non ha pretese d'interpretazione metrica. Non possiamo affermare che tra un soggetto con punteggio 6 e uno con 5 vige la distanza di "un intervallo", uguale a quella che c'è tra un soggetto con punteggio 4 e uno con 3. Ciò che si può affermare è che il soggetto con punteggio 6 è posto lungo la scala dell'atteggiamento in una posizione superiore ai soggetti con punteggio 5, 4 e 3, e via dicendo. Un esempio classico tratto da Stouffer *et al.* [1950; citato in Bruschi 1999] è quello dei sintomi della paura: vomito, nausea, accelerazione del battito cardiaco. Chi subisce il primo effetto (il più forte, quindi il più "difficile") avrà in precedenza subito anche gli altri due; chi proverà nausea non necessariamente subirà anche il vomito, ma sicuramente avrà sofferto un aumento del battito cardiaco; alcuni proveranno solo quest'ultimo; altri, infine, nessun sintomo. La logica è che se essi sono sintomi cumulativi di un attributo unidimensionale, chi ha provato un sintomo più forte dovrà necessariamente aver provato anche quelli più deboli. Questo è effettivamente il criterio deterministico più corretto e stringente per accertare l'effettiva natura unidimensionale di una serie di *item*.

3.3.2 Scalabilità e trattamento degli "errori"

Nell'ambito dello studio degli atteggiamenti e delle opinioni è però abbastanza raro trovare proprietà che rispettino in modo perfetto le condizioni di scalabilità del modello di Guttman. Alcuni studiosi (tra cui Marradi [2007]) sostengono che alla base vi sia l'indebita torsione che ha portato l'ideatore del modello a estendere le caratteristiche delle abilità e delle capacità anche al campo delle opinioni e degli atteggiamenti. In particolare sarebbe l'assunto di cumulatività a prestare il fianco alle critiche. In effetti, in un test di competenze matematiche o logiche ci si aspetta che individui capaci di risol-

vere quesiti particolarmente complessi e difficoltosi, siano a maggior ragione in grado di farlo con problemi elementari. Un'opinione, però, che costituisce l'espressione verbale di un atteggiamento, non è "facile o difficile", "superata o fallita": può essere solo condivisa o no, per cui ciò che si presta ottimamente alla sfera delle abilità, più difficilmente si adegua a quella delle opinioni. Non è un caso che il modello di Rasch stesso, che attinge abbondantemente ad assunti comuni a quello di Guttman, tra cui la cumulatività, sia nato e sviluppato nel settore dei test di abilità e solo dopo abbia trovato impiego nello studio di atteggiamenti e opinioni.

Preso comunque atto del difficile riscontro di esempi di atteggiamenti scalabili secondo criteri cumulativi, si presenta di fronte al ricercatore il limite della natura deterministica del modello. Non a caso Edwards [1957] per primo fa notare che esso è più orientato al test di un'ipotesi che non alla costruzione di una scala: esso serve principalmente a controllare se un gruppo di *item* e un insieme di soggetti rispettino congiuntamente una serie di condizioni che *definiscono* il concetto di scala cumulativa riproducibile[45]. Se ciò non accade, si dirà che una certa dimensione non è scalabile unidimensionalmente in modo cumulativo e deterministico. Questo è un esempio interessante di un modello che definisce una struttura: non si parte dalla descrizione delle caratteristiche di un certo attributo, sviluppando poi lo strumento di rilevazione in funzione del riconoscimento di quelle; si parte piuttosto selezionando una serie di assunti e costruendo lo strumento che definisce e identifica certe caratteristiche e si procede classificando le dimensioni sulla base della loro rispondenza agli assunti del modello.

Tale rigidità ha richiesto lo sviluppo di procedure per accertare quanto i dati si discostino dal modello ideale e in quali casi tale scostamento sia accettabile. Ogni volta che un soggetto risponde a un *item* in modo difforme alla previsione del modello ideale cumulativo, si parla di *errore*. Guttman stesso designa un *coefficiente di riproducibilità* che stabilisce il livello di scalabilità degli item:

$$\mathrm{CR} = 1 - \frac{\text{numero di errori}}{\text{numero totale di risposte}} \qquad \text{(eq. 3.1)}$$

[45] Sul versante probabilistico, alcuni modelli non parametrici (es. Mokken [1971]) seguono lo stesso obiettivo: controllano la scalabilità, senza pretesa di costruire anche la scala metrica.

Meno sono gli errori, più CR si avvicinerà a 1; all'aumentare degli errori, il rapporto nella parte destra della formula tenderà a 1 e CR a 0. Il coefficiente rappresenta quindi la proporzione di risposte che è possibile ricostruire, sulla base degli assunti del modello, dai punteggi dei soggetti.

	Item							errori	
soggetti	**I_1**	**I_2**	**I_3**	**I_4**	**I_5**	**I_6**	punteggi	Cornell	Goodenough
A	1	1	1	1	1	1	6	0	0
B	1	1	1	1	0	1	5	1	2
C	1	1	1	1	1	0	5	0	0
D	1	1	0	1	1	0	4	1	2
E	1	1	1	0	0	0	3	0	0
F	1	0	1	1	0	0	3	1	2
G	1	1	0	0	0	0	2	0	0
H	0	1	0	0	0	0	1	1	2
L	0	0	0	0	0	0	0	0	0
totali item	8	7	5	5	3	2			

Tab. 3.2– Ipotetico scalogramma di Guttman con calcolo degli errori sulla base di diverse procedure

Lo scalogramma della tabella 3.2 ripresenta gli stessi nove soggetti del primo esempio che rispondono a sei *item*. Abbiamo lasciato intatti i punteggi per ogni individuo, ma questa volta non siamo di fronte a un caso di perfetta riproducibilità. Come si può osservare, il soggetto B ha un punteggio di 5 ma, diversamente dalle attese, risponde positivamente all'*item* I_6 e negativamente a I_5, più "facile" del primo nei termini del modello; il soggetto D, con un punteggio di 4, rifiuta il terzo *item* e accetta il quinto; il soggetto F ha un punteggio pari a 3 ma non accetta l'*item* I_2, preferendo invece il quarto; infine, il soggetto H accetta una sola affermazione, ma non quella globalmente più "facile", come ci saremmo aspettati in caso di perfetta cumulatività, bensì l'*item* I_2. Tut-

te queste incongruenze tra gli assunti e i dati sono identificate come errori, rispetto al modello ideale. L'impatto degli errori muta in base al modo in cui sono calcolati, il che di riflesso riguarda il modo in cui sono stabiliti i punteggi stessi. Le due tecniche classiche si richiamano a due approcci diversi[46].

Il primo criterio (tecnica Cornell) definisce il numero di errori come il minimo di spostamenti che sarebbero necessari (da '1' a '0' e viceversa) per trasformare lo scalogramma osservato in uno ideale. Riprendiamo il nostro esempio. Per rendere coerente con gli assunti il profilo del soggetto B con un solo spostamento si hanno due possibilità: trasformare la risposta a I_6 da '1' a '0' (portando il punteggio a 4) o la risposta a I_5 da '0' a '1' (portando il punteggio a 6). La soluzione con meno spostamenti per il soggetto D è univoca: trasformare la risposta all'*item* I_3 da '0' a '1', modificando il punteggio in 5. E così via. Ogni spostamento produce una modifica strutturale nello scalogramma, che richiede i successivi opportuni aggiustamenti. Questa tecnica, minimizzando la stima degli errori, aumenta la soglia di scalabilità degli *item*[47]. Si noti però che la determinazione degli errori è fortemente dipendente dal campione e dagli *item* scelti, rendendo i risultati ottenuti molto legati a questi fattori contingenti. Inoltre si è potuto osservare

[46] Molte altre procedure sono state in realtà avanzate nel tempo. Successo ha avuto la proposta di Green [1954; 1956], basata su una procedura di calcolo dell'errore a più stadi e un rispettivo differente coefficiente di riproducibilità.

[47] Si noti che con questa tecnica il numero massimo di errori riscontrabili è dipendente dal punteggio, poiché essi possono essere al massimo pari al numero di risposte della categoria non modale di un certo *item* (perché cambiare le risposte della categoria modale vorrebbe dire massimizzare il numero di spostamenti, contrariamente alle istruzioni della procedura Cornell). Dati 7 soggetti, per esempio, il numero massimo di errori per un *item* è 7-*n*, dove *n* è il numero di risposte ('0' o '1') modale all'*item*, come si nota nella tabella seguente:

accettazioni dell'*item*	moda	risposte modali	max. errori
0	0	7	0
1	0	6	1
2	0	5	2
3	0	4	3
4	1	4	3
5	1	5	2
6	1	6	1
7	1	7	0

Come appare chiaramente, il numero di errori aumenta per gli *item* centrali e diminuisce agli estremi. Ciò porta a favorire automaticamente l'utilizzo di affermazioni estreme, poiché più scalabili secondo i criteri della tecnica. Per questo Menzel [1953] propone un coefficiente che prende in considerazione il numero massimo teorico di errori e non il numero totale di risposte.

che in alcuni casi la decisione sul tipo di spostamento, date alternative equivalenti, è direttamente demandata al giudizio del ricercatore. Infine, pur essendo il punteggio il fulcro della valutazione della posizione di un soggetto nel modello, la tecnica Cornell non tiene conto degli errori a partire dal punteggio effettivo. Prendiamo il soggetto F: qui minimizzare l'errore significa individuarne 1 in seconda posizione. Ma tale errore è paradossalmente riferito a un profilo con quattro risposte positive, mentre il soggetto ha un punteggio di 3.

Tale criticità è superata dal secondo criterio (tecnica Goodenough), che parte proprio dal punteggio di ogni soggetto, per calcolare la somma delle deviazioni dei profili empirici da quelli della scala perfetta [Goodenough 1944]. Nel caso del soggetto B, con punteggio 5, il profilo perfetto sarebbe 111110, per cui rispetto al profilo osservato 111101 notiamo due deviazioni, in quinta e in sesta posizione. In questo modo generalmente aumenta la stima degli errori, ma la tecnica garantisce una maggiore aderenza agli assunti del modello e al rispetto dei punteggi dei soggetti. Su questo punto è interessante notare un parallelismo con il modello di Rasch: in quest'ultimo, come si vedrà, il numero di *item* accettati, il punteggio, è la statistica sufficiente per stimare la posizione del soggetto, indipendentemente dalla combinazione di risposte che compongono i profili. La tecnica di Goodenough, molto aderente all'assunto cumulativo e al dettato della centralità degli *item* superati, si avvicina molto al punto di vista assunto dal modello stocastico: non importa come si compongono le risposte, per calcolare gli errori conta la scala ideale suggerita dal punteggio.

L'assegnazione finale dei punteggi rispecchia il tema dell'individuazione degli errori. Se si utilizza l'approccio della tecnica di Goodenough, la determinazione è univoca: si conta il numero di affermazioni accettate e si associa al soggetto. L'approccio Cornell, invece, lascia spazio maggiore all'indeterminatezza e alla scelta arbitraria del ricercatore nel decidere che punteggio assegnare nei casi in cui soluzioni diverse corrispondano a uno stesso numero (minimo) di spostamenti.

Nonostante i limiti strutturali, che abbiamo affrontato in modo necessariamente sommario e sintetico, il modello di Guttman ha avuto un grande successo nella ricerca sugli atteggiamenti, in virtù della sua impostazione rigorosa e della sua logica semplice e lineare.

A cominciare dal prossimo capitolo si entrerà nel cuore della nostra riflessione, affrontando lo studio del modello di Rasch. La vitalità della riflessione di Guttman sarà confermata dal fatto che, come vedremo, la sua impostazione è, di fatto, mantenuta nel modello logistico a un parametro del matematico danese. Quest'ultimo sviluppa una proposta che, coerente con l'approccio cumulativo, va oltre l'impostazione deterministica e ordinale, verso soluzioni stocastiche e metriche.

PARTE SECONDA

La *Item Response Theory* e il modello di Rasch

4 I principi fondamentali della *Item Response Theory*

In questa Parte Seconda affronteremo più da vicino la riflessione sul modello di Rasch. Questo è generalmente associato a una più ampia famiglia di modelli di "misurazione", raccolti sotto l'etichetta di *Item Response Theory* (IRT, Teoria della risposta all'*item*)[48]. Come abbiamo rilevato nella Parte Prima, una delle distinzioni principali nel campo dei modelli psicometrici è quella tra modelli deterministici e probabilistici. Nelle pagine precedenti abbiamo preso in considerazione tre esempi (quelli dei modelli di Thurstone, Coombs e Guttman) che ricadono nel versante dei modelli deterministici. Possiamo affermare invece che i modelli della IRT appartengono al campo probabilistico della "misurazione" e rappresentano un paradigma distinto rispetto a quello della Teoria classica dei test (TCT). Nei modelli probabilistici, quale quello di Rasch, si rifiuta l'idea di una "ordinata corrispondenza biunivoca" [Cannavò 2003] tra le dimensioni valoriali latenti, gli atteggiamenti e le opinioni che ne scaturirebbero da una parte e gli *item* che ne sarebbero l'espressione verbale dall'altra[49]. "I legami *item*-atteggiamento-valore" – prosegue Cannavò – "vengono considerati stocastici e probabilistici […], per affermare in qualche modo la temporaneità, se non la volatilità, per lo stesso insieme di attori sociali considerato, dei legami *item*-atteggiamento-valori."

Tale diversità di prospettive si sostanzia in vari aspetti. In generale, i modelli deterministici utilizzano le risposte agli *item* per costruire punteggi globali dei test: la stima dell'errore di misurazione campionario è compiuta *ex post* e parte dal presupposto che esista per ogni risposta un esito "corretto" (coerente) rispetto agli assunti del modello e uno "non corretto" (non coerente). I modelli probabilistici della IRT, invece, incorporano direttamente al loro interno una teoria dell'errore, considerano cioè la dimensione aleatoria della casualità nella loro struttura: lo scopo finale è stimare la posizione dei soggetti e degli *item* su uno stesso continuo, a partire dalla *probabilità* di ogni individuo

[48] La letteratura sull'argomento IRT e modello di Rasch è molto ampia e articolata. Per la sintetica ricostruzione generale presentata nei prossimi capitoli, si fa riferimento soprattutto a Birnbaum [1968], Lord e Novick [1968], Wright [1977a], Lord [1980], Giampaglia [1990; 2008], Hambleton, Swaminathan, Rogers [1991], Hambleton [1994], Baker [2001], Baker e Kim [2004], Cristante e Mannarini [2004], Barbaranelli e Natali [2005], Reise, Ainsworth, Haviland [2005], Maggino [2007], de Ayala [2009], Tessitore [2011].

[49] Si ricordi la distinzione terminologica di Thurstone, riportata nella sezione 3.1.

di "superare" ciascun *item*, sfruttando la struttura della teoria probabilistica che prevede tutta la gamma dei valori intermedi tra 0 (evento impossibile) e 1 (evento certo). Si tratta non solo, quindi, dell'alternativa tra una sintassi logica deterministica *versus* una probabilistica, ma anche di quella tra un *focus* sul test nella sua globalità contro un *focus* sui singoli *item* componenti.

Alle fondamenta dei modelli della famiglia della IRT si trovano i principi della Teoria del Tratto Latente (TTL), secondo la quale "la prestazione di un soggetto relativamente ad un test [può] essere predetta (o spiegata[50]) definendo le caratteristiche del soggetto stesso, legate ai tratti, stimando i punteggi dei soggetti su questi tratti e utilizzando i punteggi per predire o spiegare le *performance* sul test" [Lord e Novick, 1968]. Il tratto latente è quello non rilevabile, che non può essere "misurato" direttamente, ma che si definisce teoricamente come connesso da qualche relazione con una serie d'indicatori manifesti, direttamente operativizzabili e rilevabili, come le espressioni verbali. In conformità agli assunti teorici esplicitati nel modello, si utilizzano delle funzioni matematiche (governate o no da parametri) per descrivere le relazioni tra le proprietà osservabili e i costrutti inaccessibili al controllo diretto del ricercatore, selezionando poi quelle più coerenti con i dati raccolti.

Molti autori della IRT (tra cui Lord e lo stesso Rasch) hanno elaborato questi modelli nell'ambito degli studi sulle capacità scolastiche, associando il concetto di "tratto latente" a quello di "abilità". Il linguaggio adottato conferma tale contesto: si parla di *performance*, di superamento dell'*item*, di abilità dei soggetti e di difficoltà degli *item*. Ciononostante, si è diffusa un'interpretazione altrettanto legittima e coerente con gli assunti dei modelli IRT, che li utilizza per lo studio degli *atteggiamenti* come dimensioni latenti, interpretando l'accettazione o meno degli *item* come raggiungimento o meno di determinate soglie d'intensità dell'atteggiamento stesso, rappresentate dalle affermazioni che compongono il test[51].

Nell'ambito dei modelli parametrici della IRT, la funzione matematica identifica una curva caratteristica[52] ogivale (normale o, più frequentemente, logistica) per ogni

[50] Un termine più cauto rispetto a "predetta" e "spiegata" è quello di "riprodotta".

[51] Si tratta quindi di un meccanismo simile a quello operante nel modello di Guttman, basato sul criterio di manifestazione dei livelli di una proprietà. Si veda la sezione 3.3.

[52] Un altro insieme di modelli IRT, detti non parametrici, non ipotizza alcuna forma funzionale [Tessitore 2011].

item che, sulla base della sua posizione lungo la dimensione latente (descritta dal parametro di "difficoltà") e di eventuali altri parametri, permette di stimare la probabilità che un soggetto, condizionatamente alla propria posizione lungo il *continuum* della proprietà (descritta dal suo parametro di "abilità"), superi tale *item* (ossia concordi con esso, lo accetti, nel caso di modelli per *item* a risposta dicotomica) o comunque scelga una certa categoria di risposta (nel caso di modelli per *item* a risposta multipla). In questo senso, la combinazione delle risposte alle variabili manifeste è reinterpretata sulla base della struttura matematica probabilistica dei modelli, per arrivare infine a stimare quantitativamente la posizione degli *item* e dei soggetti lungo una scala a intervalli.

I modelli variano rispetto a una serie di fattori: il numero di parametri, di dimensioni latenti (unidimensionali o multidimensionali), di categorie di risposta (dicotomiche o politomiche). Ogni tipo di modello si fonda su una serie di assunti matematici che si riflettono sul piano teorico (e viceversa). Per esempio, scegliere un certo numero di parametri significa assumere che la funzione di risposta dipenda solo da quelli [Lord 1980]. Allo stesso modo, si assume che gli *item* selezionati per il test appartengano solamente alle dimensioni ipotizzate e non siano influenzati da altre. In questa sede, ci occuperemo solo dei modelli unidimensionali per *item* a risposta dicotomica, quale quello di Rasch: di ciò si tenga conto nell'illustrazione che segue.

4.1 Concetti preliminari

Formalizzando più organicamente quanto detto, l'assunto fondamentale della IRT è che talune proprietà siano concepibili come dimensioni quantitative continue latenti che vanno da meno a più infinito (abilità nel condurre un compito, propensione a fare qualcosa, atteggiamento verso un certo referente ecc.) e che ogni soggetto possegga un certo ammontare di tale proprietà, collocandosi idealmente in un punto preciso del *continuum*, proporzionalmente all'intensità della proprietà stessa. Quest'ultima, anche se non direttamente rilevabile, può essere connessa dalla teoria a una serie d'indicatori operativizzabili, permettendo così di rappresentare su una struttura formale (scala) quella empirica (dimensione latente). In questo modo, ogni punto del *continuum* è trasformato dalla funzione di rappresentazione in una quantità numerica della scala, che sarà a

sua volta associata ai soggetti relativamente alla loro posizione stimata sulla proprietà. Quest'ultima quantità è di solito denotata dalle lettere β o θ dell'alfabeto greco: in questo testo adotteremo convenzionalmente la prima.

Lo stesso ragionamento, *mutatis mutandis*, si applica agli *item*, che sono interpretati come gli indicatori della dimensione latente. Si assume che ognuno di essi sia espressione di un certo livello della proprietà (difficoltà di un compito, manifestazione di una certa propensione a fare qualcosa o di un atteggiamento verso un referente ecc.), che si collochi a un certo punto del *continuum* e che sia rappresentato da un certo valore sulla scala del parametro, denotato dalla lettera greca δ.

Considerato un certo *item* con una quantità δ di proprietà, ogni soggetto avrà una probabilità $P(\beta)$ di superare l'*item* (condurre a termine correttamente un compito, dichiararsi propenso a fare qualcosa, concordare con una certa intensità di un atteggiamento verso un referente ecc.). L'assunto vuole che al crescere del valore di β non debba decrescere la probabilità di superare l'*item*: la relazione tra β e $P(\beta)$ deve essere monotonica. Violare tale assunto significherebbe, per esempio, accettare l'eventualità che un soggetto più abile di un altro possa avere una minore probabilità di quest'ultimo di superare un compito connesso a tale abilità.

Riferiamoci a un caso ipotetico e si prenda come proprietà di riferimento la "tolleranza verso gli immigrati". Essa può essere pensata come una dimensione che va da un minimo (rifiuto totale verso gli immigrati) a un massimo d'intensità (apertura totale). Data una certa affermazione che esprime un determinato livello di accettazione degli immigrati, la probabilità di concordare con essa aumenta con il crescere del livello di tolleranza del soggetto rispondente. Perciò, un individuo all'estremo inferiore del *continuum* (nessuna tolleranza) avrà una probabilità molto bassa o quasi nulla di accettare una qualsiasi posizione che esprima una qualche forma di apertura verso gli immigrati, mentre un soggetto all'estremo superiore (totale tolleranza) avrà una probabilità molto alta di accettarla. Tra i due estremi troveremo tutte le possibili posizioni intermedie. Qualora i valori di β e δ coincidano, si assume che per il soggetto i due esiti (accettazione o rifiuto) siano *equiprobabili*. Tale fattispecie è alquanto importante, poiché il livello δ di "difficoltà" di un *item* è definito proprio come la quantità di proprietà necessaria affinché un soggetto abbia una probabilità pari a 0,5 di accettarlo.

La relazione tra i valori di β in ascissa e la probabilità di accettare l'*item* è descritta da una curva, cui si è già accennato, che prende il nome di *curva caratteristica dell'item* (*Item Characteristic Curve*, ICC), rappresentata solitamente da una funzione sigmoidale. Per ogni valore di δ è possibile identificare almeno una specifica ICC.

In questo contesto ci occuperemo solamente dei modelli a risposta dicotomica, ma occorre ricordare che quelli della IRT possono trattare in generale dati di tipo non cardinale, ordinali o meno[53]. Come sempre, sono considerazioni di ordine teorico a giustificare la scelta di *item* con categorie di risposta dicotomiche (*item* binari) o multiple. I primi sono spesso preferiti, anche perché garantiscono una maggiore semplicità matematica del modello. Molte volte, anche quando i dati sono raccolti con risposta libera o categorie multiple, essi sono poi ricondotti a due categorie nella successiva fase di analisi [Baker, 2001]. Nel caso di test di abilità, in cui esiste una sola risposta giusta possibile, ricondurre a due categorie (corretta/sbagliata) è del tutto intuitivo. Nel caso di *item* relativi a dimensioni attitudinali, la formulazione dicotomica è meno diretta poiché, trattandosi di significati di affermazioni verbali che utilizziamo come indicatori, difficilmente possono essere interpretati in modo talmente definito ed univoco[54] da occupare per tutti i soggetti un singolo punto dell'ipotetico *continuum*. Dobbiamo pertanto assumere che il significato di un *item* si collochi attorno a una "soglia" nel *continuum* e che solo chi si colloca al di sopra di essa avrà una probabilità maggiore di 0,5 di accettarlo.

[53] Esempi di modelli per *item* con categorie ordinate sono il *Rating Scale Model* di Andrich [1978] e il *Partial Credit Model* di Masters, con i loro relativi sviluppi. Per una panoramica ampia si rimanda a Giampaglia [2008].

[54] La vaghezza semantica è un elemento assolutamente diffuso nel linguaggio comune e non del tutto eliminabile anche in quello delle scienze sociali. Di tale alone semantico attorno a qualsiasi termine è necessario tenere conto in ogni fase di ricerca [Bruschi 1999, 2005; Giglioli e Fele 2000; Marradi 2007; Pitrone 2009].

5 Assunti e caratteristiche dei modelli IRT unidimensionali a risposta dicotomica

5.1 Gli assunti di base

Vi sono almeno tre assunti fondamentali della teoria del tratto latente comuni ai modelli IRT unidimensionali, la cui soddisfazione costituisce il prerequisito per la giustificazione teorica e matematica della loro interpretazione.

5.1.1 Unidimensionalità del tratto latente

Il primo di questi riguarda l'*unidimensionalità del tratto latente*. Questo assunto richiede che ogni *item* utilizzato per la stima del modello appartenga a una sola dimensione latente. Operativamente ciò accade se la risposta del soggetto è determinata solo dal suo grado d'intensità sulla proprietà. Se nel processo di rilevazione entrano in gioco altre dimensioni (fattori di valutazione non previsti dal modello), allora le stime di abilità e difficoltà sono distorte, non stiamo più "misurando" lungo lo stesso *continuum*, ma obliquamente attraverso *continua* molteplici. Se tutti gli *item* appartengono alla stessa dimensione e solo a quella, si può affermare che lo spazio latente è completo, cioè può essere integralmente descritto dal modello unidimensionale. Testare l'unidimensionalità di un *set* di *item* non è un'operazione semplice e univoca [Lord 1980] e, anche qualora sia accertata in modo affidabile, non risolve a un tempo anche la questione della validità degli indicatori: dimostrare che è soddisfatto il requisito di unidimensionalità è necessario per le caratteristiche matematiche del modello, ma non è sufficiente a dimostrare che la dimensione espressa dalle affermazioni sia proprio quella ipotizzata dal ricercatore.

5.1.2 Indipendenza locale

Strettamente connesso all'assunto di unidimensionalità è quello d'*indipendenza locale*: anzi, si può affermare che il primo è condizione sufficiente del secondo. L'indipendenza locale richiede che qualsiasi risposta di un soggetto a ogni *item* sia in

nessun modo influenzata dalle risposte agli altri *item* o da proprietà diverse dal tratto latente.

Da un punto di vista teorico, la condizione d'indipendenza tra più eventi si ottiene allorquando la probabilità congiunta del loro verificarsi è uguale al prodotto della probabilità del verificarsi dei singoli eventi. In questo caso è possibile ricostruire l'informazione complessa (la probabilità congiunta) come combinazione delle informazioni semplici (le probabilità singole): queste ultime riescono a riprodurre integralmente il dato della prima e sono tra loro, appunto, indipendenti. Condizione necessaria all'indipendenza è che non vi siano interazioni nelle probabilità del verificarsi dei singoli eventi.

Nell'ambito della IRT, il concetto d'indipendenza locale richiede che gli *item* siano tra loro indipendenti *condizionatamente* ai vari livelli del tratto β. In altre parole, neutralizzando l'effetto della dimensione latente nei soggetti (tenendo fermi i suoi vari livelli), le risposte agli *item* devono essere tra loro indipendenti, avere cioè le stesse distribuzioni relative. Se accade ciò, significa reciprocamente che qualsiasi rapporto di dipendenza tra gli *item* è condizionato solo dal tratto latente. Per questo si è detto che la condizione di unidimensionalità è sufficiente per quella d'indipendenza locale: se gli *item* sono espressione esclusiva di una e una sola proprietà latente, al netto di essa le risposte saranno tra loro indipendenti. Ciò equivale a dire che se lo spazio latente del modello è completo, cioè descritto in modo integrale dalle dimensioni latenti considerate (nel nostro caso una), allora vale la condizione d'indipendenza locale. Infine, si può ancora dire che se i dati si adattano a un modello di tipo unidimensionale, allora anche la condizione d'indipendenza locale sarà verificata.

Nel linguaggio formale della IRT, ciò è rappresentabile come

$$P(X = x|\beta) = \prod_{i=1}^{k} P(X_i = x_i|\beta) \qquad \text{(eq. 5.1)}$$

dove

X è la funzione di risposta sotto forma di variabile casuale;

x è la realizzazione della variabile casuale (la modalità di risposta[55]);

[55] Nel caso di specie che trattiamo in questa sede, quello dicotomico, i valori delle modalità convenzionalmente impiegati sono:

- '1' in caso di successo, ossia di *item* superato (es. il soggetto concorda con l'affermazione);

β è un livello dato di proprietà nei soggetti;

k è il numero di *item*.

L'indipendenza locale è la condizione necessaria per poter correttamente interpretare le informazioni fornite dalle risposte agli *item* ed ha quindi una valenza strategica nella struttura dei modelli della IRT.

5.1.3 Monotonicità

Il terzo assunto, già richiamato, è quello di *monotonicità*. Esso stabilisce che la probabilità di superare un qualsiasi *item* non può decrescere al crescere della quantità del tratto latente di un individuo, β. Dati due soggetti caratterizzati da β_1 e β_2, con $\beta_1 \geq \beta_2$, non può darsi per uno stesso *item* il caso in cui

$$P(X = 1|\beta_1) < P(X = 1|\beta_2). \qquad \text{(eq. 5.2)}$$

In altre parole, maggiore è l'intensità della proprietà latente nel soggetto, maggiore deve essere la sua probabilità di superare un *item* che di tale proprietà è espressione. Da un punto di vista formale, per ogni soggetto *n*, ogni item *i* e ogni valore *t*,

$$P(X_{ni} > t|\beta_n, \delta_i) \qquad \text{(eq. 5.3)}$$

è una funzione monotona della proprietà β del soggetto.

Se tutti e tre gli assunti illustrati sono soddisfatti da un modello, esso è definito come "modello undimensionale monotono localmente indipendente".

- '0' in caso d'insuccesso, ossia *item* non superato (es. il soggetto rigetta l'affermazione).

5.2 La curva caratteristica dell'*item* e il significato dei parametri

Un altro elemento ipotizzato *a priori* nella costruzione del modello (che ne caratterizza in modo fondamentale la natura), è la forma funzionale della curva che descrive la relazione tra il tratto latente e la probabilità di superare ciascun *item*. La figura 5.1 rappresenta un'ipotetica ICC (*Item Characteristic Curve*), a partire dalla quale è possibile richiamare alcune delle caratteristiche di base.

L'asse delle ascisse rappresenta la quantità di tratto latente, qui limitata a una porzione del *continuum* che va da -3 a +3, ma il cui dominio ipotetico copre tutti i possibili valori, da meno a più infinito. L'asse delle ordinate rappresenta invece la probabilità di superare l'*item* in questione e, in quanto tale, varia da 0 a 1.

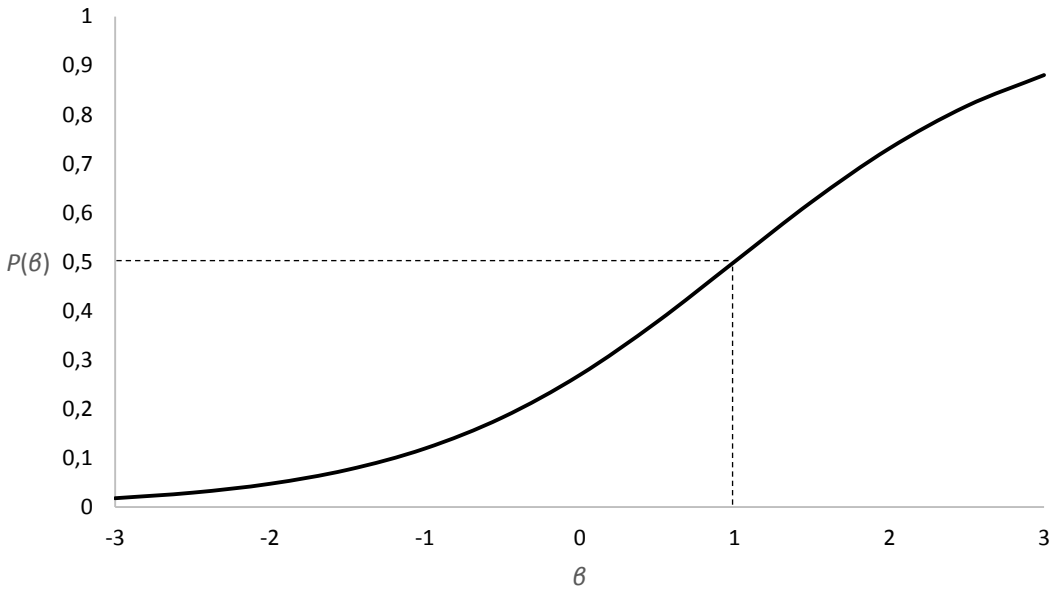


Fig. 5.1 – Esempio di curva caratteristica dell'*item*

Come si osserva, al crescere della proprietà in ascissa aumenta contestualmente la probabilità di superare l'*item*. Nel nostro esempio, i soggetti con un livello di tratto pari a 1 hanno una probabilità pari a 0,5: con livelli superiori tale probabilità tende a 1, mentre con livelli più bassi si avvicina a 0.

Dalla curva caratteristica dell'*item* è possibile desumere almeno tre caratteristiche fondamentali, che si traducono in altrettanti parametri del modello: posizione, discriminazione, *guessing*. Non in tutti i modelli a queste caratteristiche è concesso di variare e la scelta di quelle da considerare condiziona la natura della funzione in gioco. Il valore

dei parametri stimati determina le caratteristiche della ICC e, attraverso essa, identifica le proprietà strutturali imputate all'*item* in questione.

5.2.1 Curva caratteristica dell'*item*: la funzione matematica

Prima di descrivere le connessioni tra gli aspetti grafici delle curve e l'interpretazione sostantiva delle caratteristiche degli *item* dicotomici, rappresentate e riassunte dai parametri del modello, è importante specificare la *natura matematica* delle funzioni che descrivono tali curve e il posto che i parametri trovano in esse. In generale, una funzione di risposta all'*item* (*Item Response Function*, IRF) stabilisce e descrive da un punto di vista matematico la relazione tra l'abilità di un soggetto e la probabilità associata di avere come esito '1' nella scelta binaria (conoscere la risposta a un quesito, condividere un'affermazione ecc.). Una specificazione rigorosa di questa relazione da un punto di vista algebrico è un elemento indispensabile se si vuole costruire una teoria della misurazione che giustifichi i modelli e sia sottoponibile al vaglio pratico ed epistemologico: è il cuore della formalizzazione del modello stesso.

Com'è ovvio, vi è una relazione biunivoca tra la curva (rappresentazione geometrica) e la funzione che la descrive (espressione matematica). Poiché la scelta della forma funzionale avviene *a priori* (come operazione preliminare alla stima dei parametri), ci pare naturale che debba essere annoverata tra gli assunti del modello. Qualunque sia quella scelta, essa varierà di volta in volta in relazione al valore dei parametri, ma dovrà rispettare tutti gli assunti del modello (sez. 5.1).

Le funzioni utilizzate solitamente sono di due tipi: quelle derivate dalla curva normale e quelle derivate dalla famiglia delle curve logistiche cumulative. Entrambi i tipi, soprattutto nel tratto centrale, sono tra loro simili e possono essere riportati l'uno all'altro attraverso debiti fattori di trasformazione. Questa traducibilità ha permesso di considerare "intercambiabili" le funzioni, facendo sì che nella pratica si preferissero quelle matematicamente più semplici, cioè le logistiche. Le ragioni di tale esigenza acquistano consistenza se si pensa che questi modelli sono stati sviluppati ben prima della diffusione di massa di calcolatori sufficientemente potenti e dei *personal computer*, in un'epoca in cui nell'economia della ricerca era essenziale selezionare tecniche meno complesse possibili. La curva logistica, molto diffusa nello studio della crescita delle popolazioni in biologia e poi in demografia, fu derivata per la prima volta a metà del

XIX secolo. Fu alla fine degli anni ’50 del Novecento che venne introdotta nei modelli per descrivere le *item response curve*, conoscendo la diffusione citata.

In una sua forma particolare, la funzione della curva logistica cumulativa è

$$Y = \frac{1}{1+e^{-x}} \qquad \text{(eq. 5.4)}$$

dove la costante *e* è il numero di Nepero, pari a 2,718. Essa descrive una curva sigmoidale che al crescere di *x* conduce a valori crescenti e tendenti a 1, mentre mostra un asintoto tendente a 0 verso sinistra. La figura 5.2, in tutto analoga alla 5.1, illustra il tracciato di tale curva.

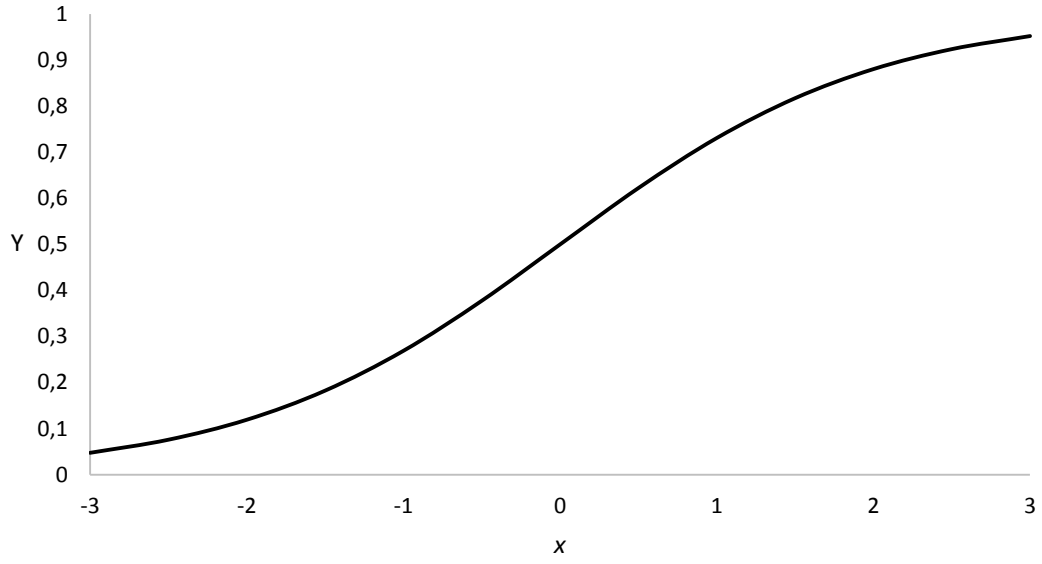


Fig. 5.2 – Curva logistica cumulativa descritta dall’equazione 5.4

I modelli logistici più diffusi per *item* dicotomici si fondano tutti sull’equazione generale appena mostrata, con le dovute varianti. Essa descrive la probabilità di superare un *item*. Speculare a essa è la curva che descrive invece la probabilità di non superare l’*item*. Tra di esse vi è un rapporto di reciprocità, i due eventi sono complementari e la loro somma è in ogni caso pari a 1: sottraendo a esso il valore di una delle due probabilità, si ricava per differenza l’altra.

Il passo successivo è capire come i parametri trovino sistemazione all’interno delle funzioni caratteristiche delle curve e quindi come esprimano il loro impatto nel modello di misurazione. I modelli della famiglia che stiamo trattando sono identificati dal

numero di parametri presi in considerazione. A questo punto possiamo dire, con maggiore precisione, che esistono modelli logistici per uno, per due o per tre parametri. Il *focus* di questa ricerca è sui modelli logistici a un parametro, noti in particolare come modello di Rasch. E' quest'ultimo in particolare su cui concentreremo l'attenzione e di cui si affronterà in modo più puntuale la logica di derivazione. Il prossimo passo è riflettere su quali siano e cosa governino i parametri dei modelli.

5.2.2 Il parametro di posizione

Il primo aspetto riguarda la posizione della curva sull'asse delle ascisse: tale caratteristica si connette direttamente al concetto di "difficoltà" o di "livello di proprietà" dell'*item*. Come si è accennato in precedenza, la difficoltà di un *item* è definita dalla quantità di proprietà latente necessaria a un soggetto per avere una probabilità pari a 0,5 di superarlo.

La figura 5.3 mostra tale parametro all'opera. La curva dell'*item* 3 (I_3) è la stessa presente in figura 5.1: perché un soggetto abbia la stessa probabilità di superarlo o non superarlo, è necessaria una quantità di proprietà pari a 1, per cui la difficoltà dell'*item* stesso sarà individuata da quel livello. La curva dell'*item* 2 (I_2) è invece spostata a sinistra rispetto alla prima: la probabilità pari a 0,5 di superarlo si verifica quando il livello del tratto latente si attesta a -1. Perciò, per avere la stessa probabilità di superare I_2 è necessaria una quantità minore di tratto latente rispetto a quella necessaria per I_3: quest'ultimo ha un livello di difficoltà maggiore. Per comprendere più a fondo il meccanismo, si noti che la curva I_2 giace completamente al di sopra di quella I_3: ciò comporta che, per qualsiasi valore di β in ascissa, la probabilità di superare I_2 sarà sempre maggiore di quella di I_3, cioè I_2 sarà sempre più "facile".

In generale, a parità di altri parametri, date due curve caratteristiche con differenti livelli di difficoltà, quella di livello più alto sarà spostata a destra e posta al di sotto dell'altra, per tutti i valori del tratto latente. Terminando l'esempio, si osservi ora che la curva dell'*item* 1 (I_1) giace a metà tra le altre e che $P(\beta) = 0{,}5$ quando $\beta = 0$: esso è quindi più "difficile" di I_2, più "facile" di I_3.

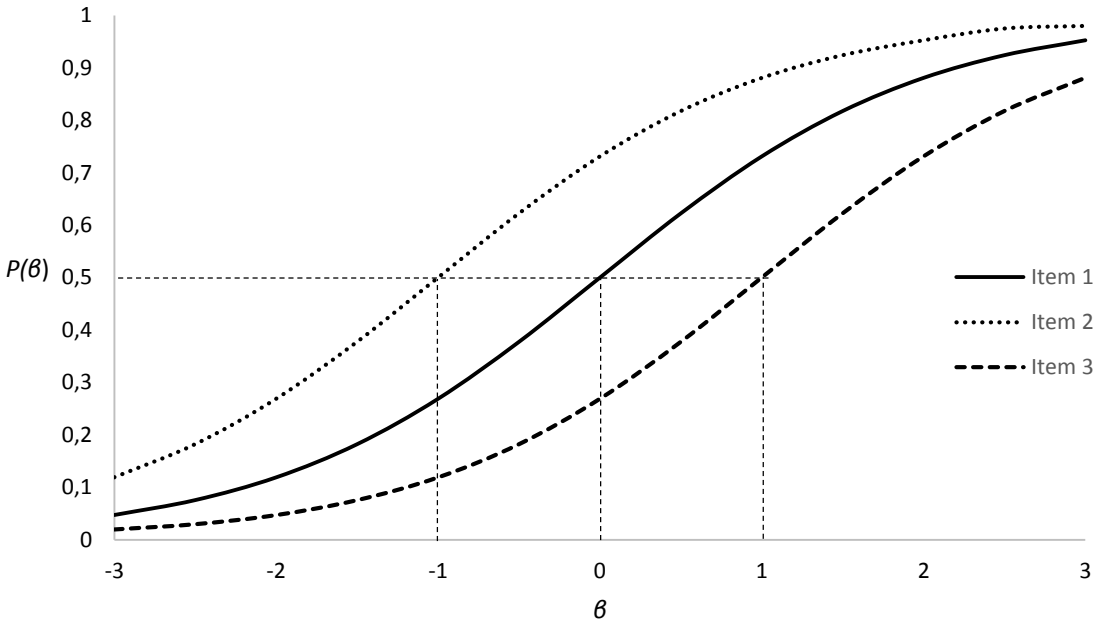


Fig. 5.3 – Raffigurazione di tre ipotetiche curve caratteristiche dell'*item* con differenti parametri di posizione

Le tre curve dell'esempio differiscono solo per i rispettivi parametri di difficoltà e sono perfettamente sovrapponibili: la crescita o la decrescita del livello di un *item* corrisponde graficamente a una traslazione della curva rispettivamente verso destra o verso sinistra. Tutti i modelli della IRT sono caratterizzati da un parametro di difficoltà dell'*item*, δ. I modelli monoparametrici, compreso quello di Rasch, considerano solo quello per descrivere la relazione tra il tratto latente e le risposte. Esso è la base concettuale per i modelli della IRT, poiché fornisce l'informazione indispensabile a stabilire un criterio di confronto tra gli individui e lo strumento di misurazione.

5.2.3 Il parametro di discriminazione

Un altro carattere della curva è la sua pendenza nella sezione centrale[56]: più essa è ripida, meglio riesce a distinguere in quella zona tra soggetti anche molto vicini tra loro; più è piatta, peggio riesce a distinguere soggetti anche abbastanza distanti. Essa è perciò un indicatore della capacità dell'*item* di discriminare i soggetti, anche se separati da piccole differenze. Tale parametro si definisce, per l'appunto, di "discriminazione". Gli ef-

[56] In particolare, si considera la pendenza massima tra quelle di ciascun punto della curva. Tale massimo si colloca esattamente in coincidenza del livello di difficoltà dell'*item*, ossia dove $P(\beta) = 0{,}5$. In tale punto, cambia anche il flesso della curva logistica.

fetti si possono notare in figura 5.4. Le tre curve hanno la stessa posizione, il che significa che possiedono lo stesso livello di difficoltà; cambia invece la loro pendenza. La curva I_1 ha la stessa di quelle delle figure precedenti. La curva I_2 mostra invece un andamento più piatto: essa giace al di sopra di I_1 fino al punto in cui $P(\beta) = 0{,}5$ per poi starne al di sotto. L'effetto è che, spostandosi lungo l'asse delle ascisse, la probabilità di superare l'*item* cresce molto lentamente: passando da -2 a +2, per esempio, la probabilità cresce di circa 0,54 contro una differenza di 0,76 nel caso della curva I_1. La conclusione è che l'*item* associato alla curva I_2 ha una minore capacità di discriminare, di distinguere due soggetti con posizioni differenti lungo la dimensione latente.

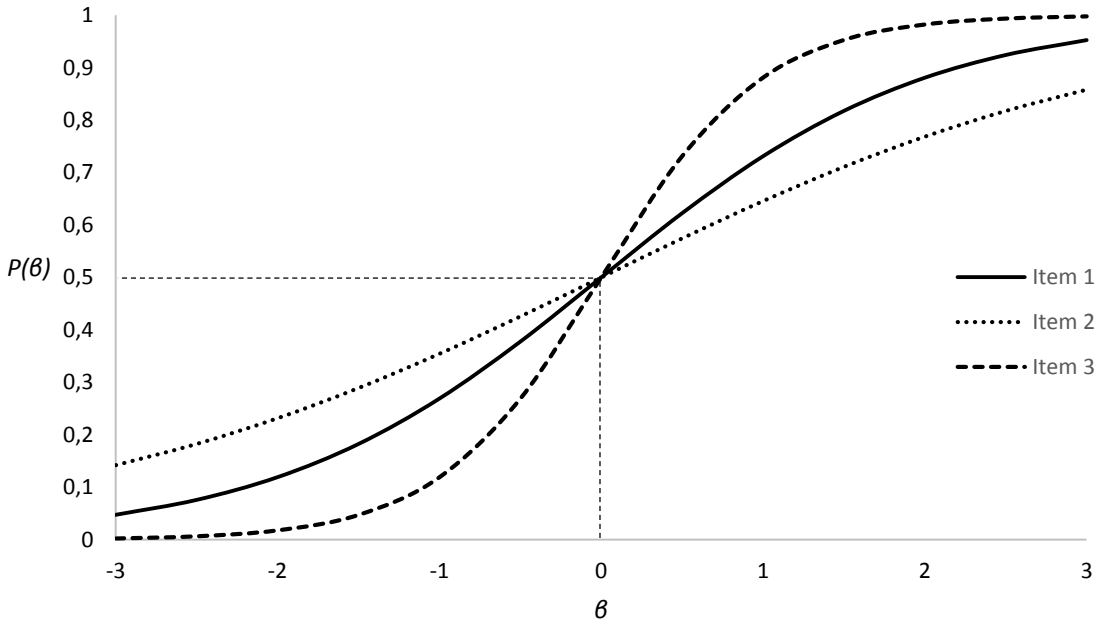


Fig. 5.4 – Raffigurazione di tre ipotetiche curve differenti tra loro per il parametro di discriminazione

L'opposto accade alla curva I_3, il cui andamento è molto più ripido delle altre due. Ciò si traduce in una più alta capacità di discriminare soggetti tra loro anche molto vicini, *in particolare nell'area centrale della curva*, quella intorno alla sua posizione δ. In questa fattispecie, passando da $\beta = -2$ a $\beta = 2$, $P(\beta)$ aumenta di 0,96. In pratica, la curva I_3 copre tutto lo spazio del codominio (0;1) in un intervallo molto ristretto del dominio.

Si osservi cosa accadrebbe nel caso-limite in cui la discriminazione assumesse un valore pari a infinito (fig. 5.5). In questa circostanza il tracciato della curva seguirebbe

quello dell'ascissa, da meno infinito fino al punto di soglia δ; da lì si dipartirebbe parallelamente all'asse delle ordinate, per poi proseguire orizzontalmente all'ascissa, con $P(\beta) = 1$, fino a più infinito. Uscendo dalla descrizione geometrica, la probabilità di superare un *item* sarebbe nulla fino al punto in cui l'abilità del soggetto è pari alla difficoltà dell'*item*; da quel punto in poi, la probabilità diventerebbe certa (cioè pari a uno). Tutti quelli con una quantità di proprietà inferiore a δ non supererebbero l'*item*, tutti gli altri sì. L'affermazione o l'oggetto in questione sarebbe quindi in grado di discriminare perfettamente i soggetti nel punto esatto del suo livello di difficoltà.

Di fatto, questo caso estremo riconduce a una fattispecie di tipo deterministico, in cui vi è un giunto rigido tra il parametro di abilità di un soggetto e quello di difficoltà di un *item*, per cui in modo sistematico il *segno* della differenza tra i due determina il superamento o meno del secondo, mentre l'intensità di tale differenza non costituisce un fattore utile a collocare i soggetti.

Se vale quanto appena detto, allora è chiaro che nella sostanza un modello perfettamente discriminante ad altro non si riduce se non al modello di Guttman nella versione dicotomica. Infatti, quest'ultimo prevede che i) una serie di affermazioni siano accettate o rifiutate da ogni soggetto in modo deterministico sulla base dell'ordine dei livelli della relativa di proprietà e che ii) i dati siano aderenti al modello se è possibile ordinare sia i soggetti sia le frasi secondo una logica cumulativa delle risposte. Nel modello probabilistico la stima della posizione è quantitativa e puntuale (quindi i soggetti e gli *item* possono essere ordinati) e, nel caso ideale di discriminazione perfetta, la risposta prevista dei soggetti è univoca e certa, determinata sulla sola base dell'ordine delle quantità relative di proprietà. Si può quindi concludere che i modelli probabilistici della IRT a due parametri sussumono il modello di Guttman come caso particolare in cui il parametro di discriminazione è infinito e la valutazione del livello di proprietà di soggetti e *item* è ordinale.

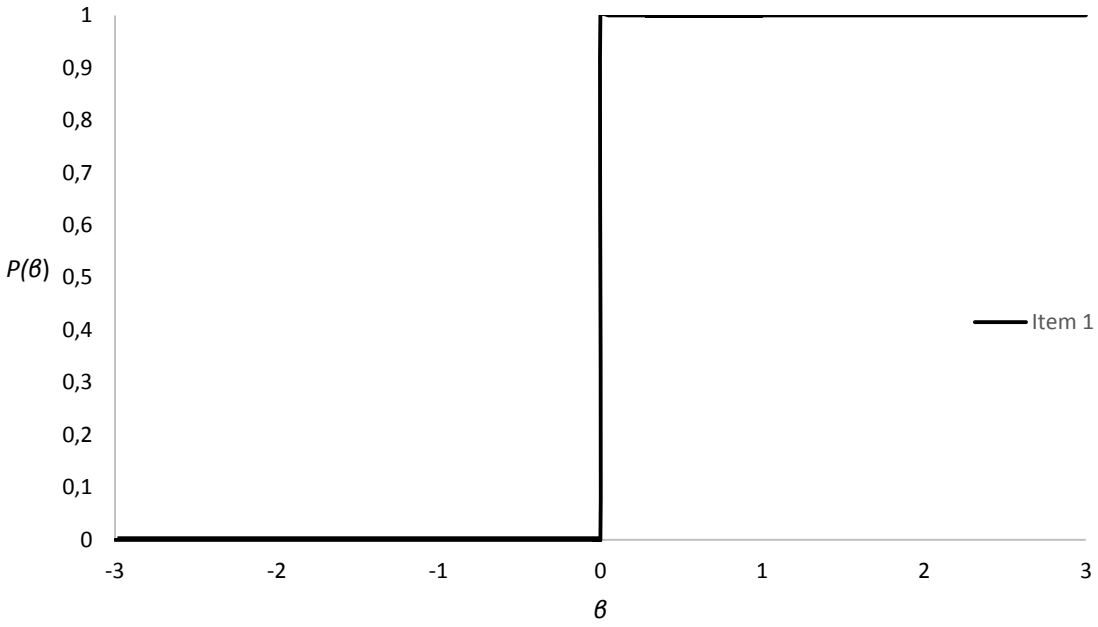


Fig. 5.5 – Caso limite di curva con potenza discriminante massima

All'opposto, se un *item* ha discriminazione pari a zero, ci troviamo di fronte al caso in cui esso non riesce per niente a discernere la posizione dei soggetti, per cui tutti avranno la stessa probabilità costante (0,5) di superarlo e la curva si ridurrà a una retta parallela all'asse delle ascisse, da meno a più infinito.

Dicendo che i modelli monoparametrici non considerano il parametro di discriminazione, s'intende che esso è mantenuto costante per tutte le ICC. Tale costante, in via di principio, è arbitraria e può essere legata alla necessità di massimo adattamento del modello ai dati. Nel caso standard del modello di Rasch essa è considerata sempre pari a 1. Si aggiunga, infine, che dagli esempi proposti s'intuisce che difficoltà e discriminazione (cioè posizione e pendenza della curva) possono variare secondo tutte le loro possibili combinazioni.

5.2.4 Il parametro di *guessing*

Finora abbiamo considerato che le curve abbiano un andamento tale per cui

$$\lim_{\beta \to -\infty} P(\beta) = 0 \qquad \text{(eq. 5.5)}$$

cioè che al decrescere della proprietà di un soggetto, la sua probabilità di superare l'*item* tenda ad annullarsi completamente. Si possono, però, dare casi in cui tale limite asintotico si attesta su valori della probabilità superiori a zero. Si pensi a cosa succede nei test di abilità o conoscenza dove vi siano domande a risposta chiusa, con un numero finito di alternative di cui solo una corretta; in questa situazione, anche chi non conosce la risposta può comunque indovinarla tirando a caso e quindi, qualsiasi sia il suo responso, egli ha sempre una probabilità diversa da zero di scegliere la risposta giusta. Di solito, utilizzando la lingua inglese, si chiama questo fenomeno *guessing effect*. E' abbastanza evidente che questo meccanismo, rappresentato da un terzo parametro nei modelli più complessi, ha un'interpretazione concettuale pienamente legittima nei test di competenza, abilità, conoscenza e cognitivi in generale; invece, nella sfera degli strumenti di rilevazione attitudinali e valoriali, non esistendo un atteggiamento o un'opinione "corretti", sfugge il senso dello stimare l'eventuale incidenza delle risposte a caso: non c'è nessuna alternativa da dover "azzeccare". Certo, un soggetto può dissimulare i propri atteggiamenti reali o le opinioni effettive, rispondendo a caso ovvero mentendo in modo ragionato sulla base di considerazioni di desiderabilità sociale, ma questa fattispecie, ci sembra, si può ricondurre nella maggior parte dei casi alla questione più generale della *validità* del dato, cioè della sua rispondenza allo stato autentico del soggetto sulla proprietà definita.

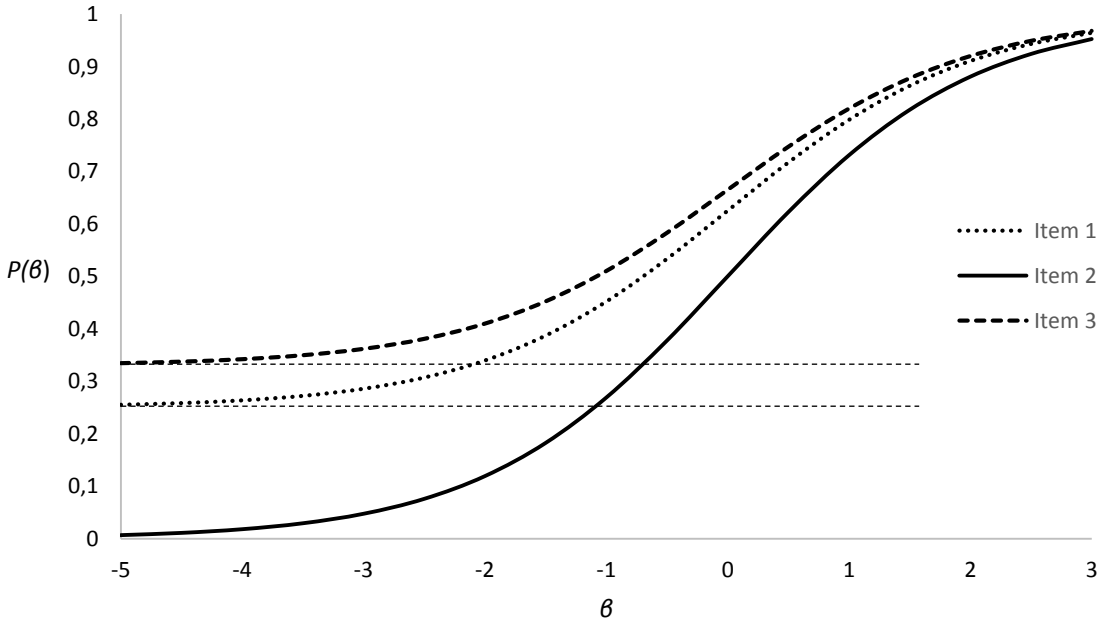


Fig. 5.6 – Rappresentazione di tre ipotetiche curve differenti tra loro per il parametro di *guessing*

Il grafico 5.6 mostra la curva I_2, con un classico asintoto a 0. Le altre due curve rappresentano invece *item* che hanno una probabilità minima di 0,25 per I_1 e di 0,33 per I_3. Se un *set* di dati si adatta a una curva come queste ultime, un soggetto con un tratto latente tendente a meno infinito avrà sempre, rispettivamente, una probabilità su quattro e una su tre di "superare" l'*item*.

5.3 La curva del test

Quanto mostrato finora ha messo in luce in modo esplicito la natura *item-oriented* dei modelli IRT: essi non si limitano a costruire stime globali dei test, ma operano stime separate del livello di difficoltà di ciascun *item* e della curva specifica che indica il livello di probabilità di superarlo per ogni possibile valore di abilità dei soggetti. Ma le batterie di rilevazione non sono mai composte di una sola affermazione. Partendo dalle curve dei singoli *item*, è possibile risalire alla stima del punteggio *vero* globale di ogni soggetto. Vi è, però, una differenza netta tra il concetto di punteggio vero nei test classici e quella della IRT: se nei primi tali punteggi si basano sulla somma *grezza* non lineare dei valori delle risposte singole, nei modelli IRT essi derivano dalla somma, per ogni livello di abilità, delle probabilità stimate di superare ciascun *item* [Baker 2001], incorporando quindi la componente aleatoria e probabilistica, secondo la formula

$$TS_s = \sum_{i=1}^{n} P_i(\beta_s) \qquad \text{(eq. 5.6)}$$

dove TS_s è il punteggio *vero* (*True Score*, in inglese) del soggetto *s*
i indica un generico *item*
n è il totale degli *item* utilizzati dall'apparato di rilevazione
P_i è la probabilità di superare l'*i*-esimo *item*
β_s è il livello di abilità del soggetto *s*

Il calcolo dei *punteggi grezzi* avviene come nei test classici, estraendo l'informazione dalla matrice dei dati che contiene i soggetti in riga, gli *item* in colonna e le risposte nelle celle. Sulla base di questi punteggi sono stimati i parametri più verosi-

miglianti del modello, dai quali a loro volta ricavare le probabilità di ogni soggetto di superare i vari *item.* La matrice-tipo dei dati, da cui tutto scaturisce, è quella rappresentata nella tabella 5.1. In riga trovano posto i soggetti *s*, che vanno da 1 a *N* (la dimensione del campione); in colonna ci sono gli *item i*, da 1 a *n* (la numerosità del *pool* di *item* somministrati al campione); nelle celle si trovano le risposte x_{si}, date dal soggetto *s*-esimo all'*i*-esimo *item*, che nel caso dicotomico di specie saranno '1' se *s* supera *i* e '0' altrimenti; i marginali di riga e di colonna rappresentano invece, rispettivamente, i punteggi grezzi r_s dei soggetti (somma dei valori delle risposte del soggetto agli *item*, $\sum_{i=1}^{n} x_{si}$) e i punteggi grezzi y_i associati agli *item* (somme delle risposte dei soggetti all'*item*, $\sum_{s=1}^{N} x_{si}$).

		item						
		1	2	3	…	…	*n*	$\sum_{i=1}^{n} x_{si}$
soggetti	1	x_{11}	x_{12}	x_{13}	…	…	x_{1n}	r_1
	2	x_{21}	x_{22}	x_{23}	…	…	x_{2n}	r_2
	3	x_{31}	x_{32}	x_{33}	…	…	x_{3n}	r_3
	…	…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…
	N	x_{N1}	x_{N2}	x_{N3}	…	…	x_{Nn}	r_N
	$\sum_{s=1}^{N} x_{si}$	y_1	y_2	y_3	…	…	y_n	

Tab. 5.1 – Matrice delle risposte dei soggetti (in riga) agli *item* (in colonna)

I punteggi r_s dei soggetti sono numeri naturali che variano da un minimo di 0 (nessun *item* superato), a un massimo teorico di *n* (tutti superati). Stesso discorso per i punteggi y_i degli *item*, che variano da 0 a *N*. I *punteggi veri*, calcolati sulla base delle curve di probabilità stimate, saranno caratterizzati dallo stesso campo di variazione, poiché anche le probabilità hanno un *range* che va da 0 a 1 per ogni *item*; trattandosi però di numeri reali, anche la loro somma sarà un numero reale. Si prendano come esem-

pio le tre curve della figura 5.3. Immaginiamo che un certo numero d'individui abbia espresso il proprio accordo o disaccordo con i tre *item* e che su questa base siano stati stimati i relativi parametri delle curve[57]. Sommando i valori delle curve per ogni punto del continuo dell'abilità, si ottiene la curva del test, mostrata in figura 5.7.

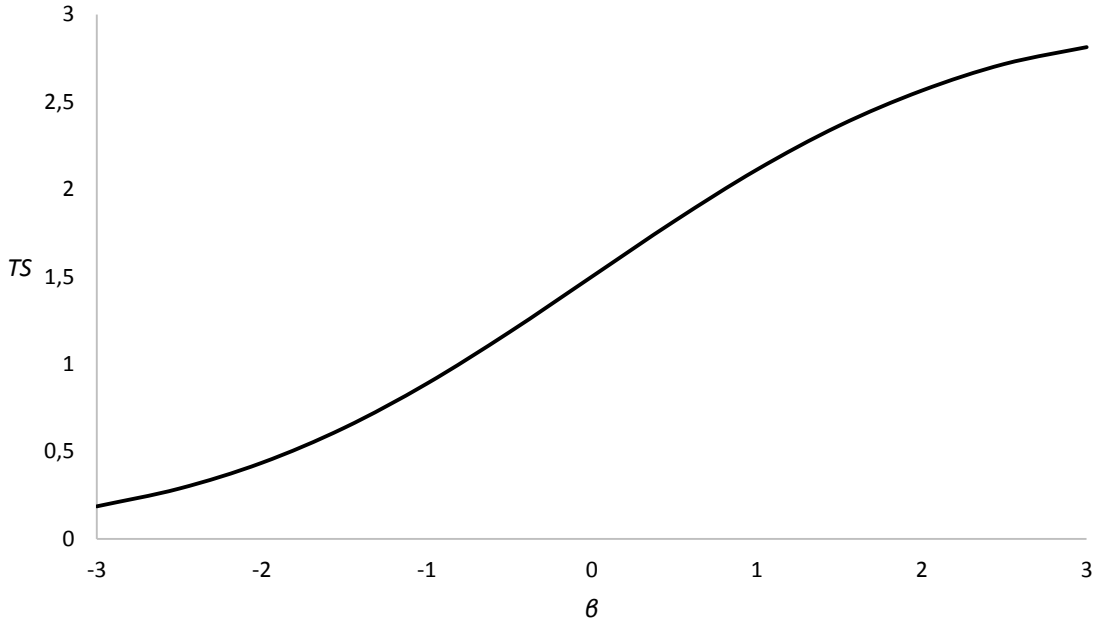


Fig. 5.7 – Curva del test per i tre *item* della figura 5.3

Essa si presenta come una curva monotonicamente crescente, con un asintoto a sinistra tendente a 0 e uno a destra tendente a $n = 3$: in generale essa non possiede una forma funzionale specifica e quindi non è caratterizzata da parametri. In questo caso, essendo la somma di tre curve isomorfiche ed equidistanti, la curva del test assume una forma simile a esse, ma questa evenienza non è necessariamente la regola. La curva descrive la relazione tra il livello di abilità e il *punteggio vero*. Per esempio, da essa si ricava che un soggetto con abilità pari a 1 ha un $TS = 2{,}11$ e un soggetto con abilità di -2 ha un $TS = 0{,}43$. Più il *TS* si avvicina a 0, più l'abilità del soggetto tenderà a meno infinito; viceversa, più il punteggio *vero* stimato si avvicina a *n*, più l'abilità tenderà a infinito.

La curva caratteristica del test, come anche quella dell'*item*, rappresenta una relazione tra due scale e quindi *non dipende* in nessun modo dalla distribuzione di frequenza delle risposte. Oltre a permettere di trasformare i livelli di abilità in *true scores*, la

[57] Non poniamoci ancora, a questo punto, il problema della bontà di adattamento dei dati al modello.

curva offre un'ulteriore interpretazione dei primi, traducendo il loro significato nei termini del *range* della scala dei punteggi. Infine occorre notare che, nonostante la curva non sia parametrizzabile e descrivibile attraverso una specifica forma funzionale, essa permette comunque di collocare lungo la dimensione della difficoltà il test nel suo insieme, esattamente come succede con le singole curve degli *item*: il livello di proprietà espresso in media dal test si individua sulle ascisse, in corrispondenza del punto della curva in cui il punteggio *vero* è di *n*/2, cioè il punto di bisezione del campo di variazione dei punteggi. In termini più semplici, la difficoltà del test è la quantità di proprietà necessaria a un soggetto per superare metà degli *item* del test stesso.

5.4 La funzione informativa

Il concetto di *informazione* in ambito statistico richiama la capacità di uno stimatore campionario di stimare con precisione il parametro reale di riferimento: maggiore è la precisione, maggiore è l'informazione che lo stimatore calcolato fornisce sul parametro incognito. La precisione è funzione inversa della variabilità delle stime attorno al parametro: a grande variabilità corrisponde bassa capacità informativa dello stimatore, mentre una variabilità vicina allo zero ne suggerisce una alta. Su queste basi, l'informazione *I* è definita come l'inverso della varianza dello stimatore:

$$I = \frac{1}{\sigma^2} \qquad \text{(eq. 5.7)}$$

Nei modelli della IRT è possibile calcolare l'informazione associata a ogni valore dello stimatore dell'abilità, $\hat{\beta}$, giungendo a una curva continua lungo tutto il suo dominio. Tali curve sono calcolate sia singolarmente per ogni *item*, sia complessivamente per l'intero test come somma delle singole funzioni informative.

In generale, la capacità informativa è massima in corrispondenza del livello di difficoltà stimato dell'*item*, mentre decresce progressivamente ai lati. Nel caso più semplice di modello, quello a un parametro[58], la funzione informativa dell'*item* è definita *in ogni punto* dal prodotto tra la rispettiva probabilità di superare l'*item* e

[58] Per l'illustrazione dei modelli di base si rimanda ai prossimi capitoli.

quella, reciproca, di non superarlo [59]. Riprendendo l'esempio della figura 5.3 e considerando un parametro di discriminazione degli *item* pari a 1, le tre curve informative sono quelle rappresentate in figura 5.8.

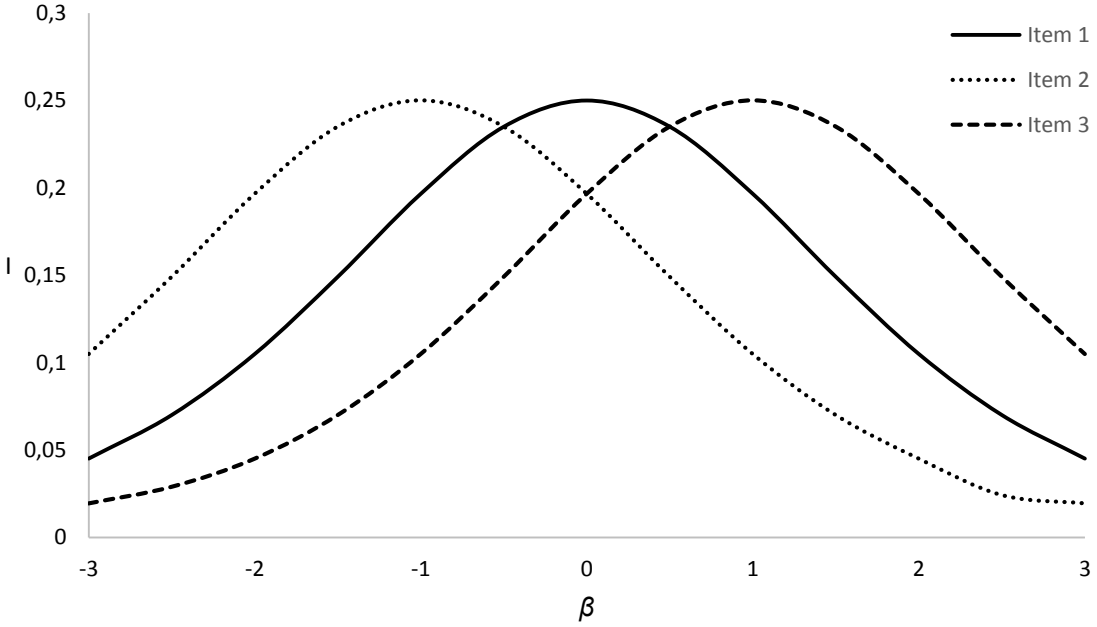


Fig. 5.8 – Funzioni informative per i tre *item* della figura 5.3

I punti di massimo corrispondono ai tre livelli di difficoltà: la stima sarà più precisa, quindi, per i soggetti la cui proprietà si attesta intorno a quei valori.

Sommando in ogni punto le funzioni informative degli *item*, si ottiene la *funzione informativa del test* (figura 5.9). In questo caso si osserva un punto di massimo per un'abilità di 0 e quindi si può concludere che il test, nel suo insieme, è più preciso nello stimare il livello di proprietà di individui collocati proprio in quel punto.

[59] Per i modelli a due parametri, tale prodotto è a sua volta moltiplicato per il valore del coefficiente di discriminazione, che nel modello standard a un parametro è uguale a 1. Per i modelli a tre parametri, il computo è decisamente più complesso, pur basandosi sugli stessi fattori. In generale, si tratta di costruire il calcolo attorno alla moltiplicazione di due serie di numeri reali a e b definiti come $a + b = c$, con c costante; per una proprietà matematica, il massimo assoluto della funzione $y = a * b$ si otterrà quando $a = b = c / 2$. In questo caso, essendo $c = 1$ (probabilità dell'evento certo) il massimo si verificherà quando $a = b = 0{,}5$. Ma le due quantità a e b sono nello specifico le probabilità di successo e insuccesso e, come sappiamo, per definizione la difficoltà di un *item* è proprio il punto del continuo in cui esse diventano equiprobabili. In questo modo si dimostra la necessità logico-matematica che la funzione informativa raggiunga il suo massimo in corrispondenza del livello di difficoltà dell'*item*.

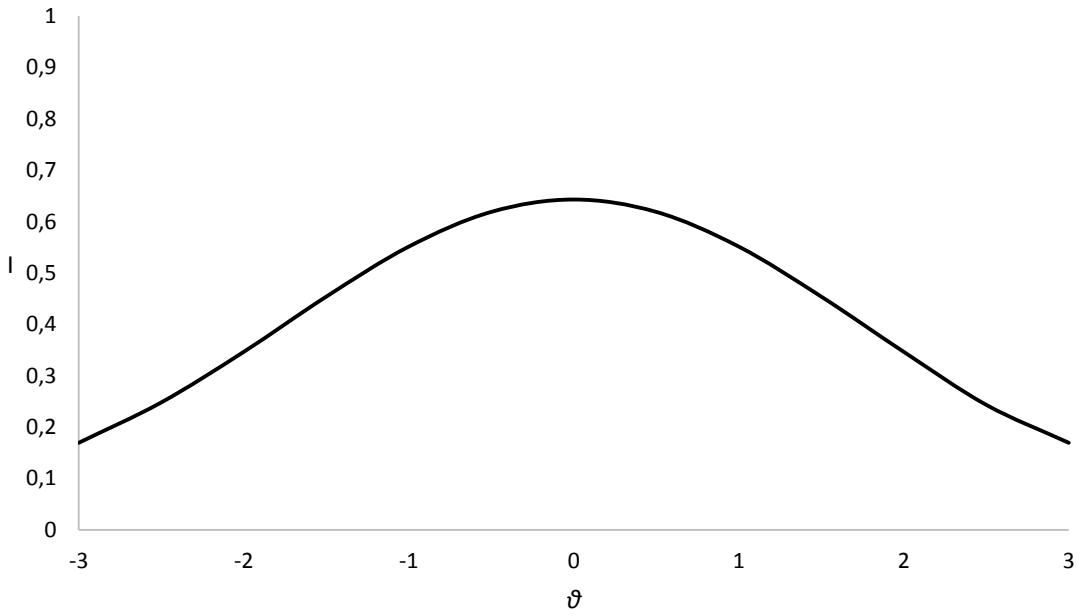


Fig. 5.9 – Funzione informativa del test di cui alla figura 5.8

In linea teorica, l'obiettivo preferibile sarebbe costruire test la cui capacità informativa si mantenga stabile lungo tutto il *continuum* della proprietà e per qualsiasi individuo, a prescindere dal suo livello di abilità. In altre parole, tale condizione garantirebbe un'eguale precisione di "misurazione" per qualsiasi soggetto, indipendentemente dalle sue caratteristiche. Ciò non è facile da ottenere nella pratica, poiché sarebbe necessario trovare una perfetta combinazione di *item*, la somma delle cui funzioni informative fosse costante in ogni punto.

Dall'altro lato, dalla mancanza di questa uniformità si possono trarre anche vantaggi. E' quanto accade quando, per motivi teorici o di ricerca, vi è l'obiettivo espresso di calibrare il test per massimizzare la precisione della stima rispetto a prestabiliti livelli di abilità. In ogni caso, le funzioni d'informazione degli *item* e del test si qualificano come strumenti molto importanti per analizzare la natura dei dati e sfruttare le potenzialità dei modelli.

6 I modelli a due e tre parametri per chiusure di risposta dicotomiche

Illustreremo adesso, in modo sintetico ed essenziale, alcuni dei principali modelli. In conformità a un criterio meramente espositivo (quindi non connesso con lo sviluppo storico dei modelli ma, anzi, in buona parte speculare rispetto a esso), partiremo illustrando la fattispecie più "complessa" (il modello a tre parametri), per poi arrivare progressivamente al modello a un parametro.

6.1 Il modello "logistico" per tre parametri

In un modello a tre parametri (3PL, 3 *Parameters Logistic*) si lasciano contemporaneamente libere di variare la difficoltà, la discriminazione e il *guessing effect*. La sua derivazione si deve a Birnbaum [1968], che partì dal modello a due parametri per sviluppare questa funzione, che in realtà si diparte dalle curve della famiglia logistica in senso stretto. Siamo nel caso incontrato in figura 5.6. Lì si era già sottolineato che l'utilizzo del terzo parametro è adatto in test di abilità a risposta multipla[60], dove con *k* categorie esiste una probabilità fissa di 1/*k* di indovinare "a caso" l'esito corretto, che si conosca o no. In questo scenario, la curva può variare per posizione, inclinazione e asintoto sinistro. La funzione è rappresentata dalla seguente equazione:

$$P_i(x_i = 1|\beta, \delta_i, a_i, c_i) = c_i + \frac{(1-c_i)}{1+e^{-a_i(\beta-\delta_i)}} \qquad \text{(eq. 6.1)}$$

dove β è il livello della proprietà nei soggetti

δ_i è il parametro di posizione dell'*item i*

a_i è il parametro di discriminazione dell'*item i*

[60] Un test di abilità avrà sempre una risposta corretta '1' e una sbagliata '0', quindi sarà sempre strutturalmente dicotomizzabile, sia che la risposta sia lasciata libera, sia che sia già dicotomizzata o prevedesse alternative multiple precostituite.

c_i è il parametro di *guessing* dell'*item i*[61]

La funzione stima la probabilità di superare l'*item*; di converso, quindi, sottraendola al valore massimo della probabilità (l'evento certo, pari a 1) si ottiene la probabilità di *non* superarlo.

Uno dei fulcri della funzione risiede nell'esponente di *e*, al denominatore. Esso contiene la differenza tra il livello di proprietà del soggetto e quello dell'*item*, il tutto moltiplicato per il parametro di discriminazione con segno negativo. Al crescere dei valori di a_i e di $\beta - \delta_i$, decresce il valore di $1 + e^{-a(\beta - \delta)}$ al denominatore e quindi cresce il valore totale della funzione. Questo, tradotto in termini sostantivi, significa che la funzione cresce (e si avvicina quindi più velocemente a 1) se aumenta la differenza tra abilità del soggetto e difficoltà dell'*item* (a vantaggio della prima), con la capacità di discriminazione che funge da acceleratore o deceleratore del meccanismo. Al contrario, diminuisce il valore della funzione e la probabilità si avvicina al minimo.

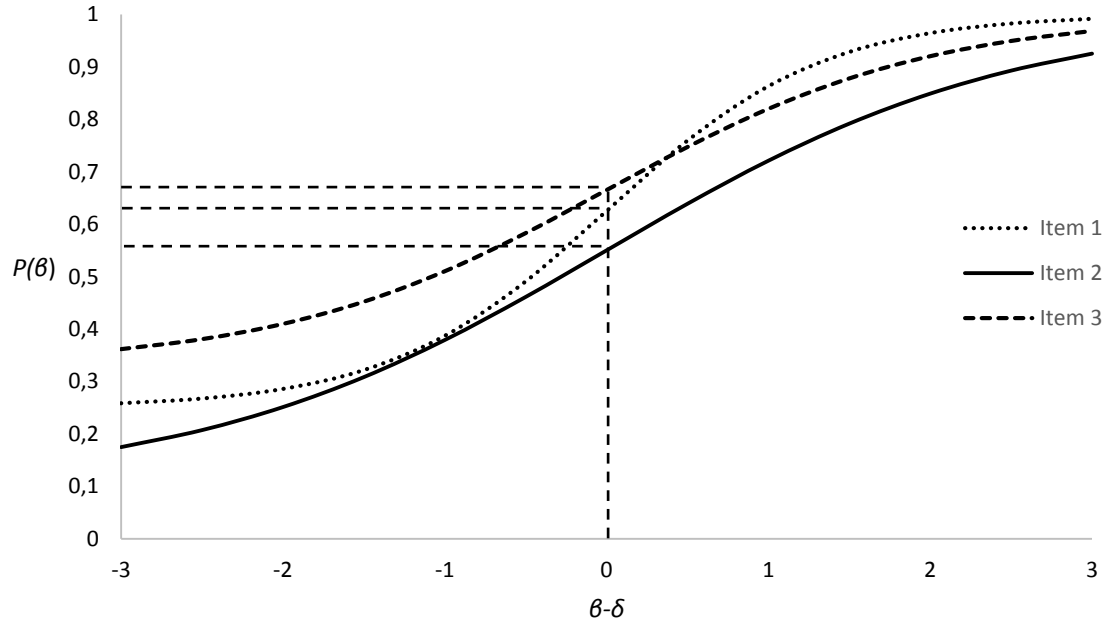


Fig. 6.1 – Modello "logistico" per tre parametri: esempi di curve caratteristiche dell'*item*

[61] Si noti, di passaggio, quella che per certi aspetti può essere interpretata come una "incongruenza" logica. Difatti, nel modello logistico a 3 parametri, il *guessing effect* è descritto da un parametro legato all'*item*, nonostante che la propensione a "tirare a caso" e indovinare la corretta modalità di risposta appartenga piuttosto ai soggetti e non alle affermazioni stesse.

La peculiarità del modello a tre parametri è poi il *guessing effect*. Il parametro c_i è pari alla probabilità di rispondere "a caso", dando la risposta corretta. Esso si presenta innanzitutto come addendo nella parte destra dell'equazione, stabilendo il valore minimo della funzione. Quando la differenza tra abilità e difficoltà tende a -∞ (l'*item* è molto più difficile dell'abilità del soggetto), il valore della funzione si avvicina sempre di più a c_i. La differenza tra 1 (il massimo teorico) e c_i (il minimo teorico) rappresenta invece il *range* di variazione della probabilità descritta dalla funzione: la sua sistemazione al numeratore della frazione serve proprio a delimitare lo spettro di valori assumibili dall'equazione.

In precedenza si è sottolineato che la difficoltà di un *item* è determinata nel punto in cui la probabilità di superarlo è 0,5; nel caso di specie dei modelli 3PL, tale affermazione si mantiene teoricamente valida, con il particolare però che, modificandosi il campo di variazione, tale punto centrale va rapportato a un minimo diverso da zero, cioè c_i. La difficoltà dell'*item* sarà, quindi, individuata dal punto in cui $P_i(x_i = 1|\beta, \delta_i, a_i, c_i) = (1 + c_i)/2$. Se, come nella figura 6.1, si usa come scala in ascissa β - δ_i, allora ogni curva sarà graficamente posta in modo da far combaciare la sua posizione nel punto 0 (in cui abilità e difficoltà si uguagliano).

Nella figura 6.1 le tre ipotetiche curve possiedono le seguenti funzioni:

$$P_1(x_1 = 1|\beta, \delta_1, a_1, c_1) = 0{,}25 + \frac{0{,}75}{1 + e^{-1{,}5(\beta - \delta_1)}}$$

$$P_2(x_2 = 1|\beta, \delta_2, a_2, c_2) = 0{,}1 + \frac{0{,}9}{1 + e^{-0{,}8(\beta - \delta_2)}}$$

$$P_3(x_3 = 1|\beta, \delta_3, a_3, c_3) = 0{,}33 + \frac{0{,}67}{1 + e^{-1(\beta - \delta_3)}}$$

Il modello 3PL è per molti aspetti un caso-limite particolare, che non possiede molte delle caratteristiche matematiche desiderabili dei modelli con uno e due parametri. Dall'altro lato, la maggiore specificazione del modello attraverso l'uso di un numero superiore di parametri, agevola la ricerca di un adattamento (*fit*) migliore ai dati: il modello è più flessibile e quindi riesce con più facilità a riprodurre le osservazioni empiriche. Studiando il modello di Rasch, vedremo in che modo quello 3PL indebolisce alcuni

assunti, fondamentali per la "misurazione" classica. Qui basti rilevare che, incorporando al suo interno il parametro c_i, contiene in se stesso un elemento del tutto estraneo alla dimensione che si sta andando a misurare.

6.2 Il modello logistico per due parametri

Quando non si considera l'evenienza di una risposta data a caso, per impossibilità pratica (assenza di alternative prestabilite) o teorica (indagini non riferite ad abilità), si può preferire un modello per due parametri (2PL, 2 *Parameters Logistic*). Da un punto di vista matematico, è come se annullassimo il valore del parametro c_i giungendo alla seguente funzione:

$$P_i(x_i = 1|\beta, \delta_i, a_i) = \frac{1}{1+e^{-a_i(\beta-\delta_i)}} \qquad \text{(eq. 6.2)}$$

dove di nuovo: β è il livello della proprietà nei soggetti
δ_i è il parametro di posizione dell'*item i*
a_i è il parametro di discriminazione dell'*item i*

In questo caso siamo in una fattispecie simile a quella presentata in figura 5.4, quando si è discusso il significato del parametro di discriminazione. Si osservi ora come cambiano le curve del modello a tre parametri, azzerando il *guessing effect* (fig. 2.11). Ovviamente, considerando la scala dell'asse delle ascisse, il punto zero è sempre quello di equivalenza tra l'intensità della proprietà del soggetto e degli *item*.

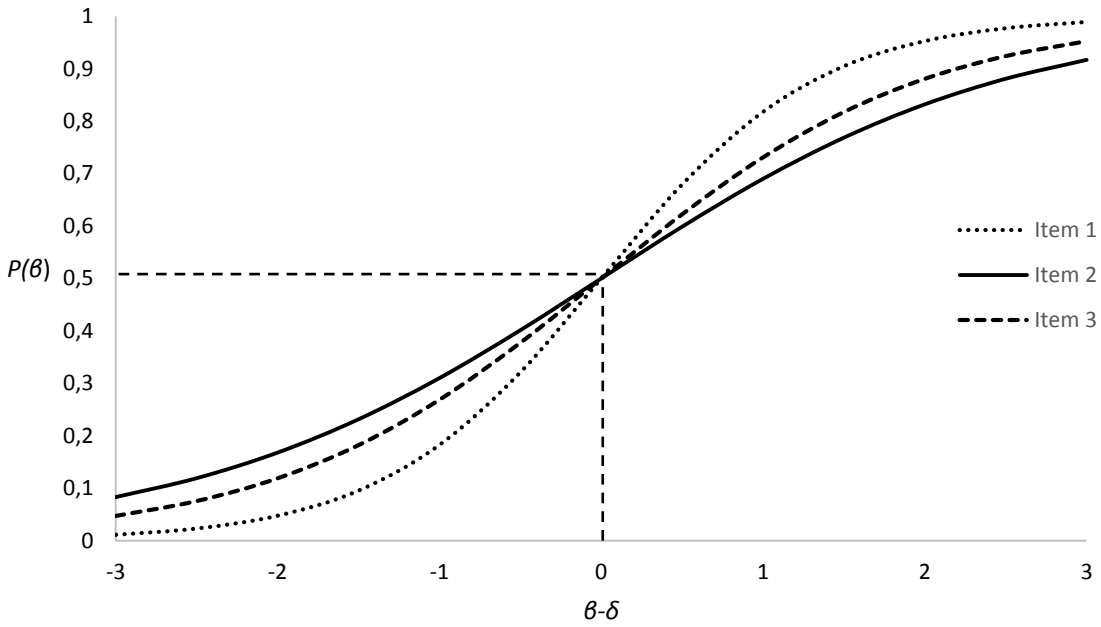


Fig. 6.2 – Modello "logistico" per due parametri: esempi di curve caratteristiche dell'*item*

Le equazioni specifiche delle tre curve del grafico sono:

$$P_1(x_1 = 1|\beta, \delta_1, a_1) = \frac{1}{1 + e^{-1{,}5(\beta - \delta_1)}}$$

$$P_2(x_2 = 1|\beta, \delta_2, a_2) = \frac{1}{1 + e^{-0{,}8(\beta - \delta_2)}}$$

$$P_3(x_3 = 1|\beta, \delta_3, a_3) = \frac{1}{1 + e^{-1(\beta - \delta_3)}}$$

In questo modello restano valide le considerazioni svolte per quello 3PL riguardo ai parametri di discriminazione e di difficoltà. Il valore di β - δ_i determina in modo monotonico quello della funzione, mentre il valore moltiplicativo di a_i si riflette nella maggiore o minore inclinazione della curva nel suo tratto centrale. Nella nostra rappresentazione grafica tutte le sigmoidi sono centrate sul punto zero poiché la scala, a differenza degli esempi portati nello studio dei parametri, rappresenta il valore dell'abilità traslato di una quantità pari alla difficoltà dell'*item*: il punto in cui tale differenza si annulla è quello in cui $P_i(x_i = 1|\beta, \delta_i, a_i) = 0{,}5$ per tutte le curve.

Rispetto ai modelli a un solo parametro, i 2PL offrono un vantaggio nell'adattamento delle curve ai dati: il parametro di discriminazione funge da "peso" moltiplicativo dello scarto tra la quantità di proprietà posseduta dai soggetti e quella espressa dall'*item*, giocando il ruolo di stimatore della sensibilità con cui quest'ultimo riesce a distinguere differenze via via più sottili tra i primi. Il servizio che il modello rende all'efficienza del *fit* comporta, tuttavia, un prezzo da pagare a livello di teoria della misurazione. Infatti, la misurazione di una dimensione cognitiva o attitudinale è sintetizzata dai punteggi grezzi (o *veri*) dei test, che comprendono le risposte a una pluralità di *item*. Tali punteggi si ottengono in questi modelli dalla somma delle singole risposte (o probabilità di risposta), sulla base degli assunti di additività e cumulatività. Questi principi possono perdere verosimiglianza nell'ambito dei modelli per due parametri. Infatti, come si osserva nelle figure 5.4 e 6.2, curve con differenti livelli di discriminazione hanno la caratteristica di non mantenere lungo tutto il dominio della proprietà la stessa relazione d'ordine con riferimento alla probabilità di superare un *item*: curve con pendenza maggiore saranno più difficili fino al punto di flesso e più facili dopo, rispetto a curve con una pendenza minore. Per fare riferimento alla figura 6.2, si ha il paradosso che un soggetto con un basso livello di proprietà troverà più facile I_2 e più difficile I_3, mentre uno con un livello alto troverà, viceversa, più facile I_3. Secondo l'abilità del soggetto, l'ordine di difficoltà degli *item* sarà quindi invertito[62] e lo strumento nel suo complesso rileverà gli stati dei soggetti in modo differenziato, mettendo in discussione che gli *item* della batteria siano tra loro cumulativi, indipendenti e unidimensionali, due assunti che dovrebbero essere fondamentali per la misurazione in ambito di IRT.

Per queste ragioni si può ritenere che il modello 2PL rappresenti un passo indietro rispetto a quello di Rasch dal punto di vista del rigore di una teoria formale della misurazione.

[62] Si può anche formulare la questione dicendo che in un modello a due parametri, l'ordine della probabilità di superare un *item* non è invariante rispetto all'abilità dei soggetti.

7 Il modello di Rasch

7.1 Derivazione e caratteri del modello logistico a un parametro di Rasch

Il modello di Rasch [1960] utilizza un solo parametro, quello di posizione.

$$P_i(x_i = 1|\beta, \delta_i) = \frac{1}{1+e^{-(\beta-\delta_i)}} \qquad \text{(eq. 7.1)}$$

Rispetto al modello 2PL, il parametro a_i "scompare", come effetto del fatto che esso è fissato al valore costante di 1. Non variando l'inclinazione delle curve, il modello di Rasch stima curve che non s'intersecano tra loro, mantenendo fisso il loro ordinamento sulla base della difficoltà, per qualsiasi valore dell'abilità di un soggetto. Se un insieme di dati si adatta in modo soddisfacente al modello, si può sostenere che gli *item* siano tra loro indipendenti e unidimensionali.

Il modello di Rasch non rappresenta semplicemente una delle varianti dei modelli della IRT, ma nasce e si sviluppa in modo autonomo, con alcune peculiarità che lo caratterizzano. Esso, con i suoi criteri molto rigidi rispetto alle curve degli *item*, è più adatto a costruire scale di misurazione fondate in modo formale, piuttosto che a fornire un adattamento descrittivo a un *set* di dati. Ricostruire il suo processo di derivazione è molto utile a comprenderne la natura.

Il ragionamento parte dalla proprietà al centro della misurazione, cioè la dimensione latente, ipotizzata come comune sia ai soggetti sia agli *item* del test. Ogni soggetto e ogni *item* sarà caratterizzato da una certa quantità di proprietà, incognita, che si vuole andare a misurare. Possiamo intanto definire le entità A_s e D_i come

A_s = quantità di proprietà posseduta dal soggetto *s* (o abilità)
D_i = quantità di proprietà espressa dall'*item i* (o difficoltà)

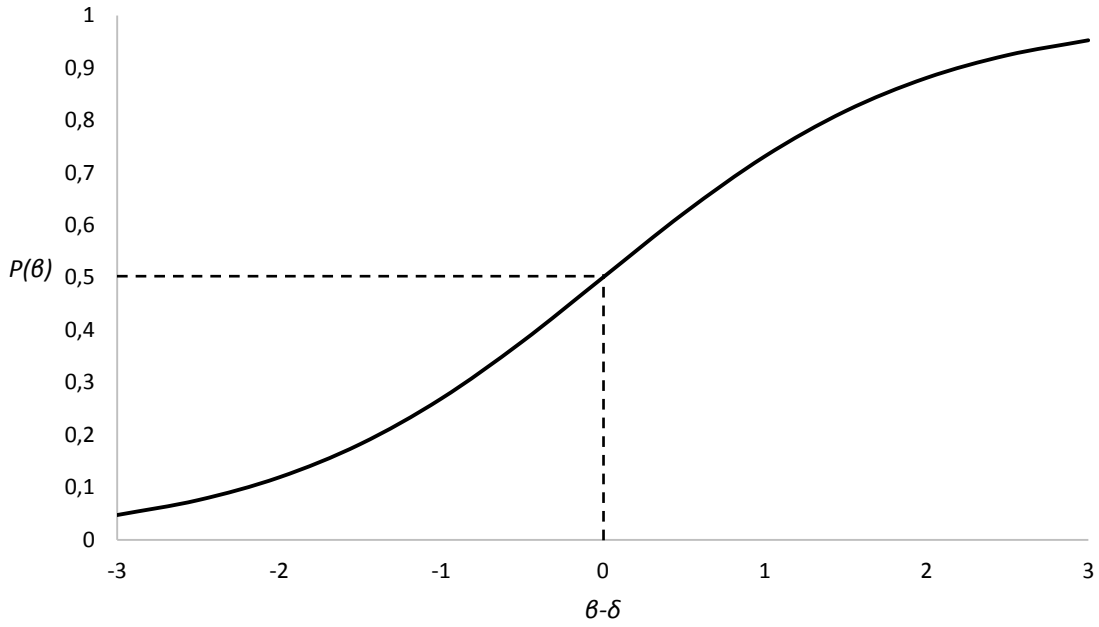


Fig. 7.1 – Modello "logistico" per un parametro (di Rasch): curva caratteristica dell'*item*

Se assumiamo che il tratto latente sia concepibile come una *quantità continua*, anche A_s e D_i saranno quantità continue, come espressioni di tale dimensione. Su questa base è ragionevole ipotizzare che il *rapporto* tra queste due quantità sia proporzionale alla probabilità di *s* di superare *i*. In termini matematici significa che[63]

$$P_i(x_i = 1|A_s, D_i) \propto A_s/D_i \qquad \text{(eq. 7.2)}$$

Via via che la quantità di proprietà del soggetto aumenta, la probabilità tende a crescere; reciprocamente, al crescere della quantità di proprietà espressa dall'*item*, tende a diminuire la probabilità di superarlo. Possiamo però andare oltre, partendo dal presupposto che $P_i(x_i = 1|A_s, D_i)$ deve variare tra 0 e 1 (per definizione) e che nel caso in cui A_s e D_i si eguaglino, i due eventi $x_i = 1$ e $x_i = 0$ diventino equiprobabili. Per cui, inserendo un fattore di normalizzazione nella formula, si ricava l'equazione per la probabilità:

$$P_i(x_i = 1|A_s, D_i) = \frac{A_s/D_i}{1+A_s/D_i} \qquad \text{(eq. 7.3)}$$

[63] Il simbolo '$\propto$' è utilizzato con il significato convenzionale di "proporzionale a".

E' facile verificare che[64]:

i) $A_s \ll D_i \implies (A_s/D_i) \simeq 0 \implies P_i(x_i = 1|A_s, D_i) \simeq 0;$

ii) $A_s \gg D_i \implies (A_s/D_i) \simeq \infty \implies P_i(x_i = 1|A_s, D_i) \simeq 1;$

iii) $A_s = D_i \implies (A_s/D_i) = 1 \implies P_i(x_i = 1|A_s, D_i) = 0{,}5.$

L'obiettivo successivo è trasformare l'espressione con le quantità A_s e D_i in una relazione di tipo lineare con la probabilità. Per farlo si sfruttano le proprietà dei logaritmi, per le quali vale l'equivalenza $x = e^{\ln x}$. Nel nostro caso, tale equivalenza diventa rispettivamente per le due quantità

$$A_s = e^{\ln A_s} \qquad \text{(eq. 7.4)}$$

$$D_i = e^{\ln D_i} \qquad \text{(eq. 7.5)}$$

A partire da queste espressioni, si definiscono i nuovi parametri di riferimento

$$\beta_s = lnA_s \qquad \text{(eq. 7.6)}$$

$$\delta_i = lnD_i \qquad \text{(eq. 7.7)}$$

che ci permettono di riscrivere le equivalenze 7.4 e 7.5 come

$$A_s = e^{\beta_s} \qquad \text{(eq. 7.8)}$$

$$D_i = e^{\delta_i} \qquad \text{(eq. 7.9)}$$

L'equazione per il calcolo della probabilità trova quindi la sua forma lineare nei termini seguenti:

$$P(X = 1|\beta, \delta_i) = \frac{e^{\beta}/e^{\delta_i}}{1+e^{\beta}/e^{\delta_i}} = \frac{e^{(\beta-\delta_i)}}{1+e^{(\beta-\delta_i)}} \qquad \text{(eq. 7.10)}$$

[64] I simboli '≪' e '≫' sono utilizzati con il significato convenzionale, rispettivamente, di "molto minore di" e "molto maggiore di"; il simbolo '≃'è utilizzato con il significato di "asintoticamente tendente a".

E' semplice dimostrare che l'ultima formulazione è del tutto equivalente a quella della formula 7.1, quindi che

$$\frac{1}{1+e^{-(\beta-\delta_i)}} = \frac{e^{(\beta-\delta_i)}}{1+e^{(\beta-\delta_i)}} \qquad \text{(eq. 7.11)}$$

rendendo assolutamente indifferente l'utilizzo dell'una o dell'altra[65].

In questo modo si rende conto della funzione logistica a un parametro di Rasch come derivazione in forma logaritmica del rapporto tra le quantità reali iniziali di proprietà del soggetto e dell'*item*. Ovviamente, per ricavare la probabilità di *non* superare l'*item* è possibile seguire lo stesso procedimento invertendo i fattori iniziali (abilità e difficoltà) oppure, molto più semplicemente, calcolare la differenza da 1 della probabilità di superare l'*item*, utilizzando i passaggi seguenti:

$$P(X = 0|\beta, \delta_i) = 1 - P(X = 1|\beta, \delta_i) \qquad \text{(eq. 7.12)}$$

$$1 - \frac{e^{(\beta-\delta_i)}}{1+e^{(\beta-\delta_i)}} = 1 - \frac{1}{1+e^{-(\beta-\delta_i)}} \qquad \text{(eq. 7.13)}$$

$$\frac{1}{1+e^{(\beta-\delta_i)}} = \frac{e^{-(\beta-\delta_i)}}{1+e^{-(\beta-\delta_i)}} \qquad \text{(eq. 7.14)}$$

[65] Si ha infatti che

$$\frac{1}{1+e^{-(\beta-\delta_i)}} = \frac{e^{(\beta-\delta_i)}}{1+e^{(\beta-\delta_i)}} \Longrightarrow \frac{1+e^{(\beta-\delta_i)}}{(1+e^{-(\beta-\delta_i)})(1+e^{(\beta-\delta_i)})} = \frac{\left(1+e^{-(\beta-\delta_i)}\right)e^{(\beta-\delta_i)}}{(1+e^{-(\beta-\delta_i)})(1+e^{(\beta-\delta_i)})} \Longrightarrow$$

$$\Longrightarrow \frac{1+e^{(\beta-\delta_i)}}{\left(1+e^{-(\beta-\delta_i)}\right)\left(1+e^{(\beta-\delta_i)}\right)} = \frac{e^{(\beta-\delta_i)-(\beta-\delta_i)}+e^{(\beta-\delta_i)}}{\left(1+e^{-(\beta-\delta_i)}\right)\left(1+e^{(\beta-\delta_i)}\right)}$$

Poiché $e^{(\beta-\delta_i)-(\beta-\delta_i)} = e^0 = 1$, i due membri diventano uguali e si dimostra l'equivalenza.

7.2 Le proprietà del modello di Rasch: discussione

7.2.1 Unicità della scala, linearità dei punteggi e unità di misura

Come si è detto, una delle caratteristiche dei modelli della IRT (e di quello di Rasch) è interpretare i livelli di proprietà di soggetti e *item* (abilità e difficoltà, rispettivamente) come parte di un unico *continuum* latente, da rappresentare su un'unica scala lineare. Inizialmente, però, le uniche informazioni possedute sulle quantità di A_s e D_i sono costituite dai punteggi grezzi del test, ossia la somma del numero di *item* superati da ciascun soggetto e la somma del numero di soggetti che supera ciascun *item*. Tale punteggio grezzo non rappresenta però una metrica adeguata e tantomeno lineare. Già nel primo capitolo, trattando dei modelli deterministici e della derivazione dei punteggi secondo diversi principi, si è osservato che essi non possono rispecchiare di per sé la struttura di una scala a intervalli, poiché le distanze relative, all'interno del *continuum*, cambiano a causa dello schiacciamento delle categorie agli estremi [Edwards 1957; Wright e Masters 1982; Marradi 2007]. Ora, tale incongruità si ripercuote anche sulle quantità di A_s e D_i, che possiamo definire come segue sulla base dei punteggi grezzi [Maggino 2007]:

$$A_s = \frac{r_s}{n - r_s} \qquad \text{(eq. 7.15)}$$

$$D_i = \frac{N - y_i}{y_i} \qquad \text{(eq. 7.16)}$$

dove: n è il numero totale di *item*
N è il numero totale di soggetti
r_s è il numero di *item* superati dal soggetto s (il punteggio grezzo)
y_i è il numero di soggetti che hanno superato l'*item* i

Come si vede, queste definizioni rispecchiano la condizione che A_s sia proporzionale in modo diretto al numero di risposte positive e in modo inverso a quello di risposte negative date da s (*odd correct*); viceversa, D_i è proporzionale in modo diretto al

numero di soggetti che rispondono negativamente e in modo inverso al numero di soggetti che rispondono positivamente all'*item i* (*odd incorrect*). Calcolando i logaritmi delle due quantità (che, ricordiamo, sono β e δ), si ottiene un doppio vantaggio. In primo luogo, essendo logaritmi di un rapporto (tra punteggi), possono essere trasformati e interpretati come differenze lineari tra logaritmi (dei punteggi). In secondo luogo, entrambe le quantità sono armonizzate e riparametrate su una stessa scala e la loro differenza si distribuisce lungo tutti i valori tra -∞ e +∞. Anche da un punto di vista concettuale, tale ridefinizione predispone il terreno per "trasformare la metrica semplice ma non lineare dei punteggi grezzi in una nuova metrica (lineare) che invece tenga conto delle distanze concettuali tra soggetti e tra *item*" [Giampaglia 2008, 48].

Il logaritmo naturale del rapporto tra A_s e D_i è uguale alla differenza dei loro logaritmi, cioè la differenza tra β e δ. Nel caso del modello di Rasch, tale differenza ha un'interpretazione ben precisa: essa è il logaritmo dell'*odd ratio* delle probabilità, ossia il rapporto tra la probabilità di superare l'*item* e quella complementare di non superarlo. Tale quantità prende il nome di *logit* per $x_i = 1$. Formalmente, quanto enunciato si dimostra così:

$$logit(X=1) = l\,n\frac{P(X=1)}{P(X=0)} = ln\frac{\frac{e^{\beta-\delta}}{1+e^{\beta-\delta}}}{\frac{1}{1+e^{\beta-\delta}}} = lne^{\beta-\delta} = \beta - \delta \qquad \text{(eq. 7.17)}$$

Se il *logit* è negativo, il soggetto ha una probabilità di superare l'*item* inferiore a 0,5; se il *logit* è positivo, la probabilità è superiore a 0,5; se il *logit* è uguale a 0, la probabilità è uguale a 0,5. Se stessimo trattando un modello di tipo deterministico, il solo segno del *logit* (in generale, il segno della differenza tra A_s e D_i) determinerebbe l'esito della risposta prevista. Il modello di Rasch, invece, come i modelli della IRT in genere, concede la possibilità che si diano esiti "anomali" rispetto a una rigida visione deterministica e quindi che esista sempre una qualche probabilità non nulla che, per esempio, un soggetto possa fallire un compito posto al di sotto delle sue capacità o, viceversa, che possa completarne uno più difficile del suo livello.

In letteratura s'individua nel *logit* l'*unità di misura* della nuova metrica ottenuta trasformando i punteggi grezzi in punteggi logaritmizzati. Senz'altro, da un punto di vi-

sta sostanziale la definizione del *logit* ha il compito di interpretare la dimensione concettuale latente rielaborata e *ancorare* la scala su cui collocare i valori stimati dei parametri di abilità e difficoltà. La proprietà rappresentata dalla scala è la "differenza tra il livello di abilità del soggetto e il livello di difficoltà dell'*item*" e il punto zero, in cui la proprietà si annulla, è il punto di equilibrio in cui il modello prevede che un soggetto abbia il 50% di probabilità di superare l'*item* descritto dalla curva. Cerchiamo di analizzare più a fondo il significato con cui si utilizza l'espressione *unità di misura* in questo caso e i motivi per cui si discosta da quello evocato nel primo capitolo a proposito della misurazione estensiva.

Assumiamo innanzitutto che i livelli di abilità e difficoltà siano effettivamente grandezze appartenenti a un medesimo attributo quantitativo. Di tale attributo si dà una definizione teorica, da cui discende un insieme di concetti operativizzabili che sono alla fine tradotti in una batteria di *item*. Il flusso teorico tra la dimensione latente e gli *item* si concretizza quindi in prima battuta in senso *top-down*. Il fatto che l'attributo quantitativo, collocato a un livello più alto nella scala di generalità, riguardi dimensioni attitudinali e sociali che risiedono nella sfera psicologica, non visibile agli individui, e non permetta di essere direttamente operativizzato, fa sì che gli indicatori stessi, a un livello più basso di generalità, svolgano una funzione connotativa supplementare rispetto alla definizione della dimensione latente. In altre parole, il modello prevede per la proprietà latente un insieme di concetti indicanti e una serie precisa di assunti sui loro legami strutturali interni e su quelli con la proprietà generale; l'eventuale non adattamento dei dati al modello stimato può voler significare che il *pool* di *item* selezionato non descrive una dimensione distinta, o che essa non è scalabile nel senso previsto dal modello. In ogni caso, questa componente *bottom-up* del flusso potrebbe portare a rivedere la definizione teorica stessa del costrutto latente non osservabile ipotizzato. Questo doppio flusso costante tra livelli diversi è ben presente nell'ambito delle scienze sociali, dove più rara è la possibilità di trovare un consenso teorico forte sulla definizione delle proprietà latenti [Bruschi 1999]. Questo aspetto è un risvolto diretto della questione (evocata nel Parte Prima) del dualismo nella misurazione tra funzioni che generano strutture e strutture che generano funzioni.

Nell'ipotesi che questo quadro delimiti verosimilmente la condizione della ricerca applicata sui costrutti attitudinali, torniamo alla questione dell'unità di misura e partia-

mo dal caso elementare in cui si confrontino un solo soggetto e un solo *item*[66] rispetto alla comune proprietà in esame, perno dell'interesse della rilevazione. Tale confronto ha almeno due caratteristiche su cui dobbiamo volgere l'attenzione. La prima è che non si richiede in partenza nessuna unità di misura *pre*-definita, ossia non c'è una data quantità di tratto latente adottata come standard di misurazione, ma solo due quantità incognite incarnate nei due poli del confronto (individuo / espressione verbale). La seconda caratteristica, condizionata dalla prima, è che la *natura* di tale confronto non riguarda la ripetizione di un'*unità* secondo una logica di concatenazione, bensì il controllo (attraverso le istruzioni della definizione operativa) di quale dei due enti esprima una quantità empirica maggiore dell'attributo definito[67]. In sostanza, questo procedimento elementare di per sé investirebbe, sì, il confronto tra due grandezze, ma non produce come esito diretto un rapporto tra esse.

Ripetendo tale operazione elementare su un insieme molteplice di *item* (da parte di una pluralità di soggetti) si giunge a una situazione analoga a quella prevista dal modello deterministico di Guttman che, introducendo un criterio di cumulatività, può controllare la sussistenza di un principio di *scalabilità ordinale*, rinunciando però affatto alla costruzione di una relazione metrica tra soggetti, tra *item* e tra soggetti e *item*. In questa prospettiva il modello *unfolding* di Coombs fa un successivo passo in avanti poiché, chiedendo agli individui di mostrare direttamente l'ordinamento di una serie di oggetti, riesce in determinate circostanze a ricavare anche informazioni sull'ordinamento delle *differenze di distanza* tra gli oggetti, controllando la sussistenza di un principio di *scalabilità metrica ordinale*.

Il modello di Rasch ambisce, invece, a esplicitare la supposta natura continua della dimensione latente su una scala a intervalli, trasformando gli esiti delle risposte agli *item* in stime di parametri sotto forma di numeri reali. La chiave di volta è costituita proprio dalla natura stocastica del modello e dall'assunto che la probabilità di superare un *item* si trovi in un determinato rapporto funzionale con la dimensione latente. In altre parole, la struttura empirica della dimensione latente (di cui le risposte agli *item* sono la manifestazione) è rappresentata sulla struttura formale della scala attraverso la media-

[66] Nella pratica della ricerca un test è ovviamente composto di una pluralità di *item*, ritenuti capaci di *saturare* nel modo più efficace possibile tutte le manifestazioni della dimensione latente.

[67] Banalmente, tale controllo potrebbe concretarsi nella domanda "Lei è d'accordo o no con la seguente affermazione?" da porre al soggetto e nella registrazione della risposta "sì/no".

zione di una terza struttura, anch'essa formale, che è la funzione caratteristica dell'*item,* interpretata secondo una logica concettuale probabilistica.

Sottolineiamo che, a questo punto della trattazione, non ci stiamo occupando del problema della stima operativa dei parametri del modello e quindi dell'adattamento tra i dati effettivamente rilevati e la curva individuata come la più verosimile: i parametri stimati associati a quest'ultima possono condurre a previsioni dei risultati anche molto diverse rispetto ai dati osservati, spingendo a rivedere o rigettare il modello. Ma noi stiamo affrontando la discussione delle proprietà formali del modello astratto, procedendo come se ci trovassimo nel caso-limite teorico in cui l'adattamento dei dati è assolutamente perfetto e, quindi, i parametri stimati perfettamente coincidenti con quelli ricavabili direttamente dalle operazioni di logaritmizzazione dei rapporti tra punteggi grezzi; in tutti gli altri casi in cui ci si allontani da questo scenario ideale, naturalmente si osserveranno dei *residui* tra il *set* di dati e il modello teorico, più o meno elevati secondo il caso.

La figura 7.2 mostra la relazione logaritmica esistente tra il rapporto delle probabilità (*odd ratio*) e le unità *logit*. I valori di $P(X = 1)$ e $P(X = 0)$, la cui somma è costante e pari a 1, stanno in una relazione di reciprocità inversa: una volta determinata la prima, la seconda seguirà per differenza. Quando la probabilità di superare un *item* è pari a 0,01 (e quella di non superarlo a 0,99), il loro rapporto è pari a $0,\overline{01}$ e il valore del *logit* si attesta intorno a -4,6. All'opposto, quando $P(X = 1) = 0{,}99$ e $P(X = 0) = 0{,}01$, l'*odd ratio* è pari a 99 e il *logit* arriva a circa 4,6. Nel caso intermedio di equiprobabilità, l'*odd ratio* delle probabilità è ovviamente pari a 1 e il *logit* si annulla. Date queste diverse circostanze, si può concludere che in situazioni reali i valori del *logit* si collocheranno quasi esclusivamente nell'intervallo tra -4 e +4.

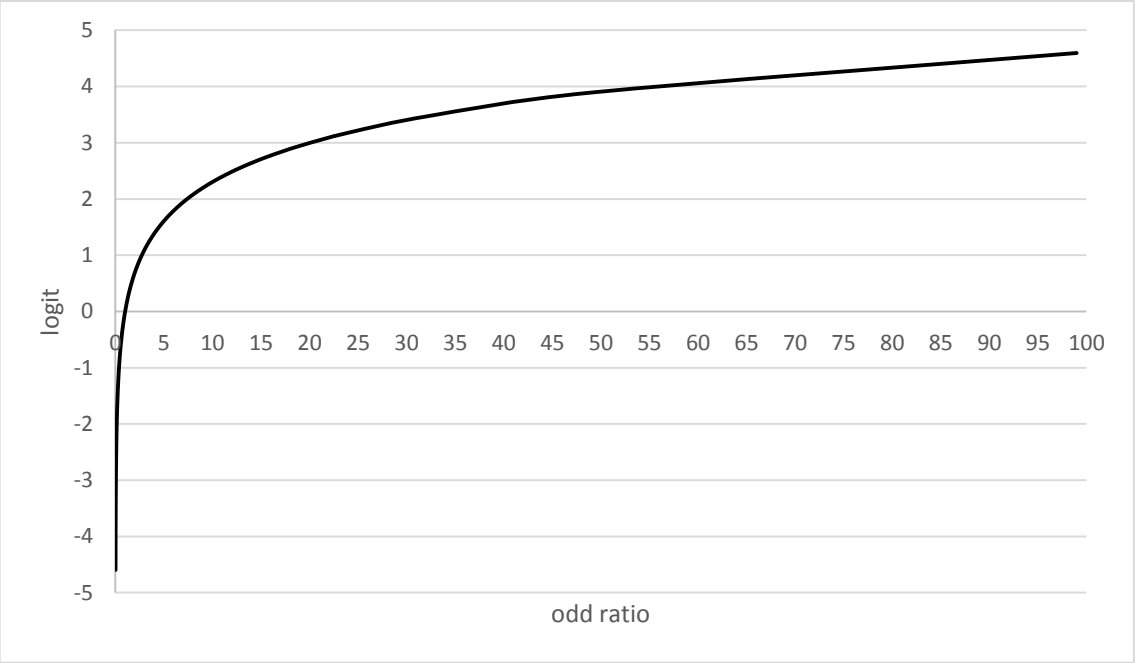


Fig. 7.2 – Curva logaritmica della relazione tra l'*odd ratios* delle probabilità e i *logit*

Quanto osservato porta a rilevare, in conclusione, che definendo il *logit* nei termini di una "unità di misura", si deve tenere presente l'uso in un'accezione diversa rispetto a quella corrente nella visione classica della misurazione. Riassumiamo alcune ragioni rilevanti, fermandoci agli aspetti concernenti la natura delle procedure di rilevazione e della struttura del modello[68].

Il primo punto è legato al tipo di definizione teorica e operativa (in termini di grandezze estensive) dei concetti di proprietà e quindi della specificazione a priori di quantità *standard*. Si pensi a uno dei casi più semplici della misurazione nelle scienze fisiche, quello della lunghezza. Un'unità di misura convenzionale di tale dimensione è il *metro*, di cui si dà una definizione operativa teorica molto precisa e fungibile (la distanza percorsa dalla luce nel vuoto in una certa frazione temporale) e che è possibile ricostruire materialmente sotto forma di oggetto che possiede la proprietà in questione e la esprime nella quantità stabilita. Nel caso della forma di misurazione adottata dal modello di Rasch la questione è diversa. Di una proprietà si può dare una connotazione generale, anche molto dettagliata e pregnante, ma non si definisce *a priori* su quella base un'unità di misura. Piuttosto, si selezionano *item* che rappresentano manifestazioni distinte della proprietà lungo il *continuum* dimensionale, ogni *item* rappresentando una

[68] Teniamo al momento ferme tutte le altre considerazioni epistemologiche riguardanti le peculiarità degli "oggetti" delle scienze umane, la loro non fungibilità ecc. [Bruschi 1999; Marradi 2007].

"soglia", che sarà superata o no da ogni soggetto, dando informazioni sulle reciproche posizioni; ma l'insieme di tali soglie, il loro ordine e la stima della loro distanza sono di fatto ignote (seppur ipotizzate) prima della "misurazione" stessa e della verifica del modello, poiché dipendenti dalle risposte degli individui, che sono a loro volta il riflesso della personale struttura interna delle opinioni esplicitate dagli *item*[69].

Il secondo punto riguarda la procedura di rilevazione: se non c'è un'unità di misura predefinita, non è possibile misurare giustapponendo un numero definito di volte tale unità, secondo una logica di concatenazione. Perciò la misura finale non sarà il frutto del rapporto tra una quantità ignota e una *standard* della stessa proprietà.

Con il terzo punto, infine, ci riferiamo al fatto che il *logit*, così come ricavato dalla stima del modello, serve piuttosto *ex post* a scalare e omogeneizzare il *continuum*, sulla base degli assunti richiesti dal modello e della struttura probabilistica, ancorando la scala al rapporto tra la quantità di proprietà del soggetto e quella dell'*item*: il punto zero è quello in cui un individuo ha la medesima probabilità di superare l'*item* e di non superarlo, nel punto esatto di uguaglianza tra ciò che è definito teoricamente come la "capacità del soggetto" e la "difficoltà dell'*item*". Queste caratteristiche della scala *logit* offrono del resto importantissime potenzialità, poiché rendono possibili tutti i procedimenti di calibrazione e confronto delle stime che, sfruttando le proprietà d'invarianza e oggettività specifica (che discuteremo in seguito), legittimano in via di principio il confronto di misurazioni diverse effettuate con campioni e *item* diversi, peculiarità molto potente e strategica nel modello di Rasch. Da questo punto di vista, la funzione del *logit* che riteniamo più decisiva non è tanto quella di *unità di misura* in senso classico quanto di *unità di scala e calibrazione*.

7.2.2 Le statistiche sufficienti

Una delle proprietà più importanti del modello di Rasch è quella delle statistiche sufficienti, a loro volta premessa necessaria per le proprietà d'invarianza del modello. Il

[69] Si prenda una dimensione latente attitudinale, per esempio quella del "conservatorismo", indicata da un insieme di *item*. Si può anche stabilire un punto di riferimento, per esempio la "quantità di conservatorismo necessaria a superare in media un certo *item i*", scelto magari perché particolarmente significativo da un punto di vista teorico; ma di per sé non si possono rappresentare direttamente altri *item* come espressione di *n* volte la quantità di conservatorismo di *i*. Per questo il concetto più appropriato è quello di *soglie ordinate* e non di unità ripetibili.

concetto di *statistica sufficiente*, così come elaborato a partire dalla riflessione di Ronald A. Fisher in poi, riguarda la capacità di una funzione campionaria di rappresentare in modo sintetico un'informazione, senza perderne la ricchezza. In altre parole, "la statistica sufficiente rappresenta una riduzione dei dati che conserva l'informazione contenuta nei dati" [Andersen 1977, 80; cit. in Giampaglia 2008].

Nel presentare gli assunti dei modelli della IRT, abbiamo sottolineato l'importanza di quello d'indipendenza locale e la sua stretta connessione con quello di unidimensionalità. Riassumendo quei concetti, la logica del modello di Rasch è semplice: gli unici fattori che influenzano le risposte degli individui sono i parametri considerati dalla funzione di risposta all'*item* e, al netto di quelli, le risposte devono essere tra loro indipendenti. Se tale assunto è rispettato, la conseguenza necessaria è che la probabilità di una certa combinazione di risposte sarà uguale al prodotto della combinazione delle probabilità singole. Per esempio, dato un certo soggetto *s* e il suo profilo di risposte a *n item*[70], sfruttando le proprietà degli esponenziali, si avrà che:

$$P(x_{s1}, x_{s2}, x_{s3} \dots x_{sn}) = P(x_{s1})P(x_{s2})P(x_{s3}) \dots P(x_{sn}) = \quad \text{(eq. 7.18)}$$

$$= \frac{e^{x_{s1}(\beta_s - \delta_1)}}{1+e^{(\beta_s - \delta_1)}} \frac{e^{x_{s2}(\beta_s - \delta_2)}}{1+e^{(\beta_s - \delta_2)}} \frac{e^{x_{s3}(\beta_s - \delta_3)}}{1+e^{(\beta_s - \delta_3)}} \dots \frac{e^{x_{sn}(\beta_s - \delta_n)}}{1+e^{(\beta_s - \delta_n)}} = \quad \text{(eq. 7.19)}$$

$$= \frac{e^{[(x_{s1}+x_{s2}+x_{s3}+\dots+x_{sn})\beta_s - x_{s1}\delta_1 - x_{s2}\delta_2 - x_{s3}\delta_3 - \dots - x_{sn}\delta_n]}}{\prod_{i=1}^{n}(1+e^{(\beta_s - \delta_i)})} \quad \text{(eq. 7.20)}$$

Poiché la sequenza $x_{s1}, x_{s2}, x_{s3}, \dots, x_{sn}$ altro non è se non la serie di numeri '1' e '0' che costituiscono la combinazione di risposte[71] del soggetto *s*, allora la somma dei suoi componenti sarà il numero di *item* superati dal soggetto, ossia il suo punteggio grezzo r_s:

[70] I profili di risposta dei soggetti corrispondono ai vettori-riga nella matrice riportata in tabella 5.1: ogni cella contiene la singola risposta x_{si} all'*item* in colonna, mentre il marginale r_s rappresenta il punteggio grezzo.

[71] Ricordiamo che, convenzionalmente, il codice '1' è assegnato alle risposte positive, cioè agli *item* superati; il codice '0' è assegnato alle risposte negative, cioè agli *item* non superati.

$$(x_{s1} + x_{s2} + x_{s3} + \cdots + x_{sn}) = r_s \qquad \text{(eq. 7.21)}$$

Per cui, è possibile riscrivere l'equazione 7.20 come segue:

$$P(x_{s1}, x_{s2}, x_{s3} \ldots x_{sn}) = \frac{e^{(r_s\beta_s - \sum_{i=1}^{n} x_{si}\delta_i)}}{\prod_{i=1}^{n}(1+e^{(\beta_s - \delta_i)})} \qquad \text{(eq. 7.22)}$$

Come mostra l'equazione 7.22, r_s funge da coefficiente di β_s, cioè dell'abilità del soggetto, ed è indifferente al tipo di combinazione che lo produce: quel punteggio grezzo è ottenuto parimenti da molte diverse sequenze di risposte. La probabilità di ottenere, in generale, una qualsiasi combinazione di risposte la cui somma sia pari a r_s è [Giampaglia 2008]:

$$P(x_{s1} + x_{s2} + x_{s3} + \cdots + x_{sn} = r_s) = \sum_{(x)|r} \frac{e^{(r_s\beta_s - \sum_{i=1}^{n} x_{si}\delta_i)}}{\prod_{i=1}^{n}(1+e^{(\beta_s - \delta_i)})} \qquad \text{(eq. 7.23)}$$

dove il simbolo $\sum_{(x)|r}$ indica la sommatoria di tutti i profili di risposta che corrispondono a un dato punteggio r. Orbene, conoscendo il punteggio grezzo del soggetto ci si può chiedere quale sia la probabilità di averlo ottenuto attraverso una certa specifica combinazione di risposte. Da un punto di vista matematico ciò equivale a calcolare il rapporto tra la probabilità di ottenere quella data serie di risposte e la probabilità di ottenere qualsiasi combinazione di risposte la cui somma sia r_s, ossia il rapporto tra l'equazione 7.22 e la 7.23. Tale probabilità condizionata al punteggio risulta:

$$P(x_{s1}, x_{s2}, x_{s3} \ldots x_{sn}|r_s) = \frac{\frac{e^{(r_s\beta_s - \sum_{i=1}^{n} x_{si}\delta_i)}}{\prod_{i=1}^{n}(1+e^{(\beta_s - \delta_i)})}}{\sum_{(x)|r} \frac{e^{(r_s\beta_s - \sum_{i=1}^{n} x_{si}\delta_i)}}{\prod_{i=1}^{n}(1+e^{(\beta_s - \delta_i)})}} = \qquad \text{(eq. 7.24)}$$

$$= \frac{e^{(-\sum_{i=1}^{n} x_{si}\delta_i)}}{\sum_{(x)|r} e^{(-\sum_{i=1}^{n} x_{si}\delta_i)}} \qquad \text{(eq. 7.25)}$$

La semplificazione della eq. 7.24, che conduce alla eq. 7.25, rappresenta un risultato teorico assolutamente capitale per il modello di Rasch. L'elemento immediatamente evidente è la scomparsa del parametro di abilità dalla formula. Formalmente questo implica che, dato un certo punteggio, il parametro β non rientra nel calcolo della probabilità che il punteggio stesso sia ottenuto attraverso una qualche specifica sequenza di risposte, ma contano solo le difficoltà degli *item*. Nella sostanza, se il soggetto supera un certo numero *r* di *item*, il profilo di risposte che ha condotto a quel risultato non fornisce informazioni in più sulla quantità di proprietà posseduta dal soggetto stesso. Da ciò deriva l'importantissima conseguenza che r_s è una statistica *sufficiente*, di per sé, per la posizione dei soggetti sul *continuum*.

Vi sono altre due conseguenze cruciali della formula 7.25 [Giampaglia 2008, 53]. La prima è che l'assenza del parametro β rende teoricamente ininfluente la distribuzione dell'abilità tra i soggetti del campione su cui si opera la rilevazione. In altre parole, dal punto di vista formale del modello la stima dei parametri δ di difficoltà degli *item* è indipendente dall'abilità degli specifici soggetti inseriti nel campione (in inglese, *sample-free*). Torneremo nel prossimo paragrafo su questo punto in modo più diffuso, allorché si tratterà della questione dell'invarianza. Basti qui osservare che questo è un aspetto di profonda differenza rispetto alla teoria classica dei test, i cui modelli *test-oriented* sono dipendenti dalle caratteristiche dei soggetti. Nel caso del modello di Rasch, *item-oriented*, il punteggio totale è una statistica sufficiente per la stima di parametri specifici per ogni *item*, che in condizioni di aderenza dei dati al modello non risentono dei caratteri di ogni specifico campione.

La seconda, ulteriore conseguenza è che in questo contesto qualsiasi distribuzione di β nel campione dei soggetti è ugualmente ammissibile, per cui, a differenza di molti altri modelli statistici, non sono necessari altri assunti sulla sua forma funzionale[72]. Questo carattere del modello è senz'altro desiderabile in un ambito, come quello della ricerca sociale, in cui le assunzioni sulle forme funzionali delle distribuzioni campiona-

[72] Sono molto diffusi i casi in cui, ad esempio, è necessario assumere che determinati caratteri si distribuiscano nella popolazione secondo la curva normale o qualche altra funzione.

rie sono raramente confermate dai dati, pur essendo in via di principio necessarie rispetto ai modelli statistici utilizzati[73].

L'indipendenza della stima dei parametri δ da quelli β trova nel modello di Rasch un riscontro simmetrico nell'indipendenza della stima dei parametri β da quelli δ [Giampaglia 2008, 53]. Vediamo come ciò sia giustificato matematicamente. Si considerino le equazioni da 7.18 a 7.20, ma stavolta si prendano come riferimenti non i vettori di riga (i profili di risposte dei soggetti) bensì quelli di colonna, cioè i profili di risposta agli *item*. Per ogni *item i*, in condizione d'indipendenza locale, varrà che la probabilità del verificarsi di una certa combinazione di risposte è uguale al prodotto delle probabilità delle singole risposte, come segue:

$$P(x_{1i}, x_{2i}, x_{3i} \dots x_{Ni}) = P(x_{1i})P(x_{2i})P(x_{3i}) \dots P(x_{Ni}) = \quad \text{(eq. 7.26)}$$

$$= \frac{e^{x_{1i}(\beta_1 - \delta_i)}}{1+e^{(\beta_1 - \delta_i)}} \frac{e^{x_{2i}(\beta_2 - \delta_i)}}{1+e^{(\beta_2 - \delta_i)}} \frac{e^{x_{3i}(\beta_3 - \delta_i)}}{1+e^{(\beta_3 - \delta_i)}} \dots \frac{e^{x_{Ni}(\beta_N - \delta_i)}}{1+e^{(\beta_N - \delta_i)}} = \quad \text{(eq. 7.27)}$$

$$= \frac{e^{[-(x_{i1}+x_{2i}+x_{3i}+\cdots+x_{Ni})\delta_i + x_{1i}\beta_1 + x_{2i}\beta_2 \mp \beta_3 + \cdots + x_{Ni}\beta_N]}}{\prod_{s=1}^{N}(1+e^{(\beta_s - \delta_i)})} \quad \text{(eq. 7.28)}$$

Similmente a quanto visto per i soggetti, la sequenza $x_{1i}, x_{2i}, x_{3i}, \dots, x_{Ni}$ è la serie di risposte (sotto forma di cifre '1' e '0') all'*item i*. La loro somma sarà il numero di soggetti che superano l'*item*, dato da y_i:

$$(x_{1i} + x_{2i} + x_{3i} + \cdots + x_{Ni}) = y_i \quad \text{(eq. 7.29)}$$

L'equazione 7.28 si riscrive quindi così:

$$P(x_{1i}, x_{2i}, x_{3i} \dots x_{Ni}) = \frac{e^{(-y_i\delta_i + \sum_{s=1}^{N} x_{si}\beta_s)}}{\prod_{s=1}^{N}(1+e^{(\beta_s - \delta_i)})} \quad \text{(eq. 7.30)}$$

[73] Tanto è vero che alcuni autori, consapevoli di tale condizione, propongono tecniche specifiche per il trattamento dei dati in modo tale da avvicinarli alle condizioni richieste. Si veda, per esempio, la tecnica di deflazione di Marradi [2007].

In questo caso abbiamo $-y_i$ che funge da coefficiente di δ_i, cioè della difficoltà dell'*item*. Come r_s, anche y_i è indifferente al tipo di combinazione da cui scaturisce. La probabilità di ottenere una tra tutte le possibili combinazioni di risposte la cui somma sia a y_i è:

$$P(x_{1i} + x_{2i} + x_{3i} + \cdots + x_{Ni} = y_i) = \sum_{(x)|y} \frac{e^{(-y_i\delta_i + \sum_{s=1}^{N} x_{si}\beta_s)}}{\prod_{s=1}^{N}(1+e^{(\beta_s - \delta_i)})} \qquad \text{(eq. 7.31)}$$

dove $\sum_{(x)|y}$ è la sommatoria di tutti i vettori-colonna che corrispondono al punteggio totale di y_i. Sulla base del punteggio finale dell'*item* si può, anche in questo caso, calcolare la probabilità di ottenerlo attraverso una combinazione data di risposte, dividendo la probabilità di ottenere quella precisa serie di risposte con la probabilità di ottenerne una qualunque la cui somma sia y_i. Essa è pari a:

$$P(x_{1i}, x_{2i}, x_{3i} \ldots x_{Ni}|y_i) = \frac{\frac{e^{(-y_i\delta_i + \sum_{s=1}^{N} x_{si}\beta_s)}}{\prod_{s=1}^{N}(1+e^{(\beta_s - \delta_i)})}}{\sum_{(x)|y} \frac{e^{(-y_i\delta_i + \sum_{s=1}^{N} x_{si}\beta_s)}}{\prod_{s=1}^{N}(1+e^{(\beta_s - \delta_i)})}} = \qquad \text{(eq. 7.32)}$$

$$= \frac{e^{(\sum_{s=1}^{N} x_{si}\beta_s)}}{\sum_{(x)|y} e^{(\sum_{s=1}^{N} x_{si}\beta_s)}} \qquad \text{(eq. 7.33)}$$

E' così che si arriva all'altro importante risultato del modello, simmetrico rispetto a quanto ottenuto partendo dai vettori dei soggetti. Questa volta a scomparire dalla formula è il parametro di difficoltà per cui, dato un certo punteggio dell'*item*, il parametro δ non è coinvolto nel calcolo della probabilità che il punteggio sia stato ottenuto attraverso una data sequenza di risposte. Se un *item* è superato da un certo numero *y* di soggetti, il risultato non è influenzato dal livello di proprietà dell'*item* stesso, bensì dall'abilità dei soggetti. Su questa base si conclude che anche y_i è una statistica *sufficiente*, in questo caso rispetto alla posizione degli *item* sul *continuum*.

Dalla formula 7.33 possiamo far discendere due implicazioni che, di fatto, sono analoghe a quelle individuate per il parametro β. L'assenza del parametro δ nell'equazione ne rende indifferente la distribuzione nel *pool* di *item* selezionati. For-

malmente, la stima dei parametri β di abilità dei soggetti è indipendente dalla difficoltà degli *item* (in inglese, *item-free*). Anche su tale punto torneremo nel paragrafo successivo, perché esso costituisce una caratteristica fondamentale del modello, nota come *oggettività specifica*, che sancisce in via di principio la possibilità di produrre le medesime stime dell'abilità dello stesso soggetto, indipendentemente dagli *item* utilizzati (e quindi delle loro difficoltà relative). Che si usino affermazioni più o meno facili, il modello offre gli strumenti tecnici e logici per giungere alla stessa stima: un elemento, questo, che si allinea alle prerogative delle procedure tipiche della teoria classica della misurazione.

L'altra implicazione è che qualsiasi distribuzione di δ nel *pool* di *item* è accettabile, senza vincoli di forma.

Concludiamo questo paragrafo con due importanti considerazioni, connesse a quanto si è appena illustrato. La prima riguarda le conseguenze delle statistiche sufficienti sulla stima dei parametri [Maggino 2007; Giampaglia 2008]. I procedimenti di stima più diffusi, utilizzando l'assunto d'indipendenza locale e procedure reiterative, si basano sulla massimizzazione della funzione di verosimiglianza (ML, *Maximum Likelihood*) per calcolare i parametri degli *item* che meglio descrivono i dati a disposizione[74]. Vi sono fattispecie che stimano contemporaneamente i parametri e altre che invece, sfruttando la loro indipendenza, li calcolano separatamente. La funzione di verosimiglianza (L, *Likelihood*) di base per il modello di Rasch, nel caso più generale di stima congiunta, è

$$L([x_{si}]|\boldsymbol{\beta}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \prod_{s=1}^{N} P_{si}^{x_{si}} (1 - P)_{si}^{x_{si}} \qquad \text{(eq. 7.34)}$$

in cui: $[x_{si}]$ è la matrice delle risposte degli N soggetti agli n *item*
$\boldsymbol{\beta}$ e $\boldsymbol{\delta}$ sono i vettori dei parametri di abilità e difficoltà
P_{si} è la probabilità che il soggetto s superi l'*item* i

In ogni caso, le procedure sono incapaci di stimare la posizione di soggetti o di *item* i cui vettori in matrice siano composti di valori costanti (solo '1' o solo '0' in cia-

[74] Più precisamente, i parametri che, partendo dai dati ottenuti, stimano la curva che con maggior probabilità può averli prodotti.

scuna cella), ossia i soggetti che hanno superato tutti gli *item* ovvero nessuno e gli *item* che sono stati superati da tutti i soggetti ovvero da nessuno. Per questo motivo, tali vettori devono essere eliminati dalla matrice e non considerati per la stima. Questa limitazione porta a una perdita d'informazione, la cui entità varia chiaramente da situazione a situazione. Concettualmente, però, questo *handicap* ha una sua interpretazione del tutto legittima. Infatti, un soggetto o un *item* i cui vettori siano composti di tutti valori '1' saranno, rispettivamente, estremamente abile e difficile: ma quanto abile e quanto difficile? Quale termine di paragone relativo si può utilizzare per quantificare questa presenza estrema di proprietà? La risposta è: nessun termine di paragone. Allo stesso modo, un soggetto o un *item* i cui vettori siano composti solo dal codice '0' saranno, rispettivamente, estremamente poco abile e facile: quanto poco abile e quanto facile? Anche in questo caso nessun termine di paragone esterno viene in soccorso per fissare una possibile stima. Per questo motivo è necessario decurtare le righe e le colonne con tali caratteristiche.

La seconda considerazione riguarda invece il confronto tra il modello di Rasch e quello di Guttman a proposito del ruolo svolto dai punteggi grezzi. Si è già osservato che il primo può essere interpretato come variante stocastica di un modello cumulativo deterministico. Adesso è interessante notare un parallelo indicativo. Nel modello di Rasch, in caso di perfetto adattamento dei dati agli assunti del modello, il punteggio grezzo di un soggetto e quello di un *item* sono statistiche sufficienti per stimare i parametri β e δ, cioè le loro posizioni lungo la scala comune. Nel modello di Guttman, in caso di perfetta aderenza dei dati agli assunti del modello si è nella condizione di perfetta riproducibilità della scala, cioè il punteggio grezzo fornisce una previsione precisa della posizione ordinale del soggetto. Insomma, in entrambi i modelli, nel caso-limite ideale, la somma delle risposte contiene tutto il succo dell'informazione necessaria a collocare soggetti e affermazioni lungo la proprietà. Nel secondo caso, il punteggio fornisce in più anche l'informazione relativa a *quali item* sono stati superati e da *quali soggetti*, in un'ottica rigidamente deterministica. Il vettore empirico di risposte non aggiunge invece informazioni statistiche nel modello di Rasch, ma c'è da rilevare che in quest'ultimo il

fulcro della stima risiede nei parametri, per i quali il punteggio è una statistica sufficiente[75].

7.2.3 Invarianza e oggettività specifica

Il paragrafo precedente ha mostrato un risultato importante del modello di Rasch[76], ossia l'indipendenza dei parametri dei soggetti e degli *item* rispettivamente dagli *item* e dal campione dei soggetti. Questa caratteristica ci introduce a un'importante proprietà (forse la più importante in assoluto) del modello, cioè l'*invarianza* dei parametri. Infatti, se è vero che questi ultimi sono *item-* e *sample-free*, allora segue che il parametro di un *item* dovrà essere lo stesso a prescindere dal campione scelto e il parametro di un soggetto lo stesso a prescindere dagli *item* ai quali risponde. In particolare, quest'ultima caratteristica è nota come *oggettività specifica*.

Come già accennato nel corso del testo, qui ci troviamo di fronte a una differenza assolutamente costitutiva rispetto ai test classici in ambito psico- e sociometrico. Infatti, laddove la stima è rappresentata dai punteggi grezzi globali, il livello medio della proprietà espresso dai soggetti del campione e dagli *item* della batteria influisce direttamente sul risultato finale. Dato lo stesso campione e due batterie di affermazioni relative allo stesso *continuum* latente, una delle quali composta di *item* facili e una da *item* difficili, si otterranno verosimilmente due risultati diversi: il punteggio medio ottenuto sulla prima batteria sarà superiore a quello sulla seconda, *pur trattandosi di stime della stessa dimensione latente sullo stesso campione*. Allo stesso modo, data una stessa batteria di affermazioni sottoposta a due campioni distinti, uno formato da individui più abili e uno da individui meno abili, si otterrà prevedibilmente un punteggio medio maggiore nel primo campione e inferiore nel secondo.

Nel caso del modello di Rasch invece, come documentato nel paragrafo precedente, i parametri che si vanno stimando sono indipendenti dalla loro relativa distribuzione tra i soggetti e tra gli *item*: infatti, essi sono messi in relazione, su un *continuum* infinito, soltanto con la probabilità di superare un *item* e il fatto che un a un certo livello di abili-

[75] Chiaramente, anche nel modello di Rasch, il livello di precisione della stima (l'informazione) e la verosimiglianza saranno maggiori in termini assoluti nel caso in cui *item* e soggetti rispettino più strettamente il principio di cumulatività di stampo deterministico [de Ayala 2009].

[76] In parte comune anche ad altri modelli della IRT precedentemente illustrati. In questa occasione, concentrandoci sul modello di Rasch, non allargheremo agli altri la riflessione su questo punto.

tà corrispondano o no individui (e quanti) è del tutto indifferente. In realtà, come cercheremo di ribadire, la situazione è più complessa: a stretto rigor di termini *la proprietà di invarianza appartiene ai parametri del modello e non alle sue stime* [Lord 1980; Hambleton 1994] ed è il grado in cui esso si adatta ai dati specifici in analisi che indica quanto la proprietà sia verosimile nei singoli casi di ricerca.

In generale, comunque, l'invarianza dei parametri è considerata parte necessaria di un processo di "misurazione fondamentale"[77]. Il concetto di *oggettività specifica* nel modello di Rasch, così come definito da Wright [1968], comporta che

> primo, la calibrazione degli strumenti di misurazione deve essere indipendente dagli oggetti che sono utilizzati per la calibrazione. Secondo, la misurazione degli oggetti deve essere indipendente dagli strumenti che sono utilizzati per misurare.

L'importanza di ciò[78] è affermata in termini più generali da Wright e Stone [1999], forse non senza un eccesso di enfasi[79]:

> Il progresso della scienza dipende dall'invenzione, dalla costruzione e dal mantenimento di misure utili. La scienza vive sulla misurazione. La misurazione vive sull'oggettività. Un termine comune per oggettività è generalità. L'oggettività è l'aspettativa e, quindi, il requisito che la quantità e il significato di una misura siano stati separati bene a sufficienza dallo strumento di misurazione e dalla situazione della misurazione, cosicché la misura possa essere usata come una quantità, sen-

[77] Wright parla di "misurazione fondamentale", utilizzando un'accezione più ampia rispetto a quella di Campbell: non si tratta, infatti, di una misurazione di entità concatenabili, bensì più in generale additive. La Parte Terza sarà dedicata all'esame del rapporto tra il modello di Rasch e la teoria assiomatica della misurazione congiunta additiva, che si presenta come un modo alternativo di dimostrare la natura quantitativa della proprietà rilevata. Poiché, come si vedrà, quella teoria è stata associata a misurazioni fisiche di tipo derivato, nel senso di Campbell, è bene non fare confusione tra queste connotazioni diverse espresse con la medesima etichetta da autori diversi.

[78] Si ricordi anche quanto discusso nella Parte Prima riguardo alle concezioni del ruolo della misurazione nella ricerca scientifica.

[79] Un'enfasi che finisce quasi per essere un po' riduttiva verso qualsiasi impresa scientifica e umana che non si ponga l'obiettivo della quantificazione.

> za riserve per quanto riguarda quale fosse lo strumento particolare o quale fosse la particolare situazione.
> Sebbene un'occasione di misurazione sia necessaria per dare origine a una misura, l'utilità della misura dipende dalla scomparsa dalla considerazione delle specificità della situazione. Deve essere possibile prendere la situazione come data e, per un momento, dimenticarsi di essa. Non fosse possibile una tale separazione del significato dalle circostanze della sua situazione, non solo la scienza ma anche il commercio, e perfino la comunicazione, diventerebbero impossibili.

Non è chiaramente questo il contesto per addentrarsi nelle possibili conseguenze nefaste della mancanza di oggettività nel commercio e nelle comunicazioni, per cui ci limiteremo a descrivere i motivi per cui il modello di Rasch può garantire, in via di principio, misure (stime) che non siano dipendenti dalla situazione e dallo strumento della misurazione[80].

7.2.4 La logica dell'invarianza nella stima degli *item*

Nel paragrafo sulle statistiche sufficienti, si è mostrato che i parametri di abilità non forniscono informazioni ulteriori rispetto al punteggio grezzo dei soggetti nella stima della difficoltà di un *item*, poiché scompaiono dalla formula della probabilità condizionata di ottenere una certa combinazione di risposte dato un certo punteggio. Si è visto che vale anche il *vice versa* ed è così che si presenta l'invarianza: i punteggi grezzi sono statistiche sufficienti per la stima dei parametri e abilità dei soggetti e difficoltà degli *item* non influenzano a vicenda, nel modello, i relativi parametri.

Si può ulteriormente illustrare la logica di funzionamento dell'invarianza nel modello mostrandola graficamente [Baker 2001]. Immaginiamo che una serie di soggetti abbia risposto in modo dicotomico a una serie di *item*. Nel grafico di figura 7.3 i punti rappresentano i gruppi di casi con pari abilità e la curva è quella che meglio si adatta ai

[80] Per un'affascinante ricostruzione storica e teorica del concetto di 'oggettività' nelle scienze umane e dei molteplici significati che ha assunto nel tempo e tra correnti di pensiero, si rimanda al bel libro di Montuschi [2006]

dati nel descrivere il rapporto tra l'abilità dei soggetti e la probabilità di superare l'*item* in questione[81].

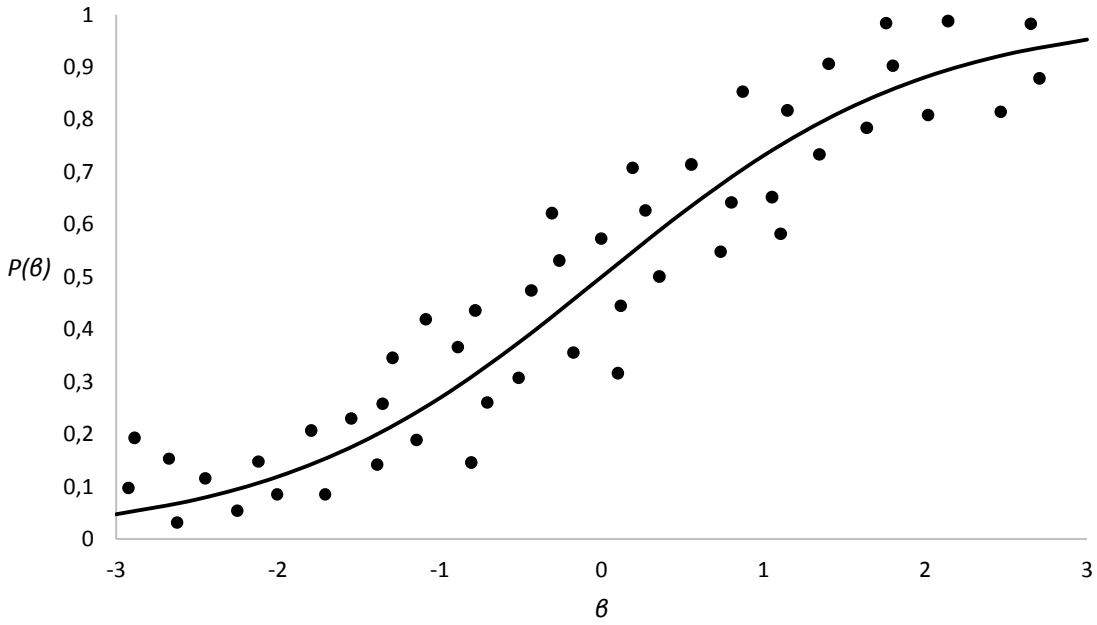


Fig. 7.3 – Ipotetica serie di dati e relativa curva stimata, rappresentati graficamente

Poiché da un punto di vista matematico la stima è indifferente all'abilità dei soggetti e quindi alla distribuzione delle caratteristiche dei soggetti in una qualsiasi popolazione, in linea teorica si otterranno gli stessi risultati anche ripetendo il procedimento di stima su sotto-gruppi di abilità diversa. Si pensi quindi di suddividere il continuo del tratto latente in quattro categorie: inferiore a -1,5; compreso tra -1,5 e 0; compreso tra 0 e +1,5; superiore a +1,5. Per ogni classe s'immagini ora di stimare i parametri dell'*item* con la relativa curva. I risultati per ogni classe saranno quelli rappresentati nelle figure da 7.4 a 7.7.

[81] Trattandosi di un esempio di fantasia, siamo stati liberi di immaginare dati molto ben adattabili a una curva logistica. Si osservi che, per quanto riguarda i punti, ciò che si descrive in ordinata è la proporzione di soggetti con una data abilità in ascissa che hanno superato l'*item*, poiché a partire da tale proporzione empirica è costruita la curva.

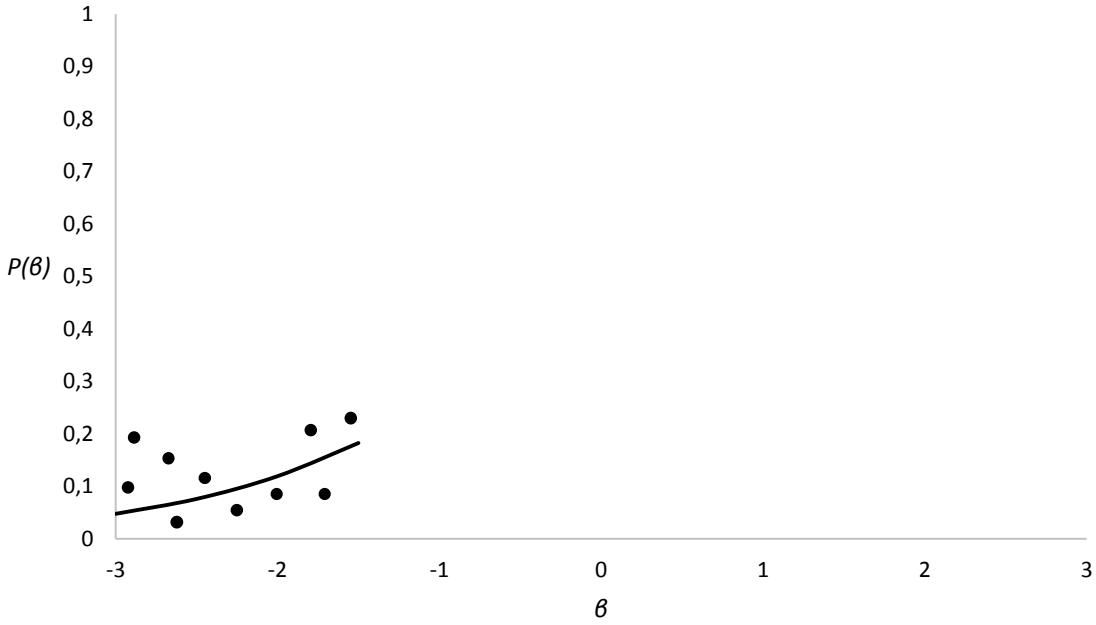


Fig. 7.4 – Prima classe: gruppi di soggetti di abilità inferiore a -1,5.

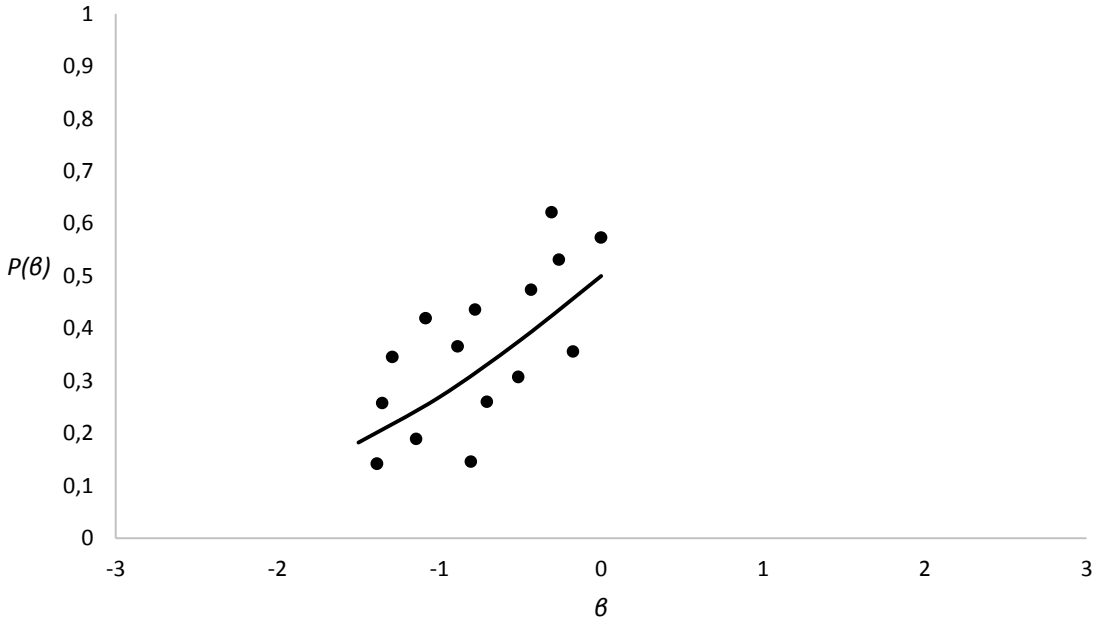


Fig. 7.5 – Seconda classe: gruppi di soggetti di abilità compresa tra -1,5 e 0.

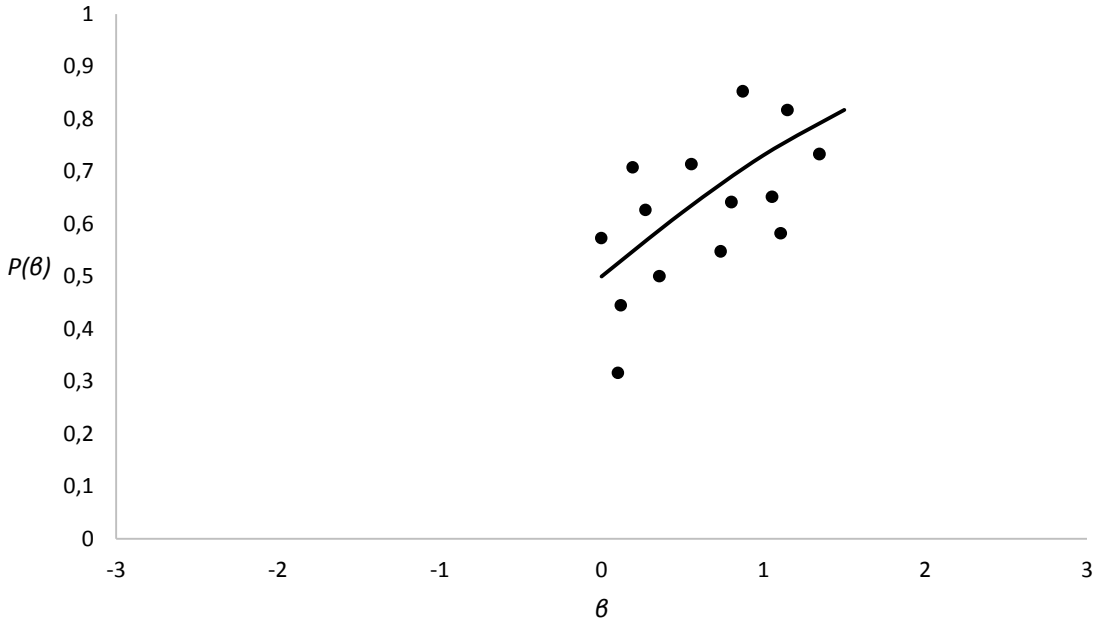


Fig. 7.6 – Terza classe: gruppi di soggetti di abilità compresa tra 0 e +1,5.

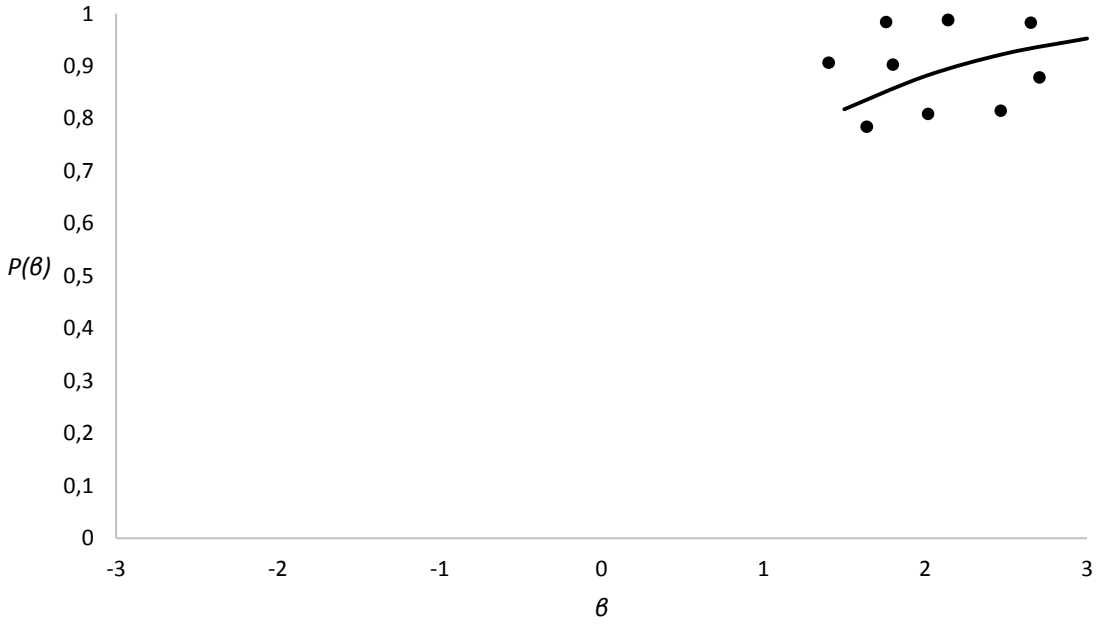


Fig. 7.7 – Quarta categoria: gruppi di soggetti di abilità superiore a +1,5.

Nei grafici sono rappresentati i tratti di curva relativi al *range* di abilità di ogni sotto-gruppo, ma questo non significa che la stima riguardi solo quel segmento: il nostro è solo uno stratagemma grafico, perché i parametri ogni volta stimati identificano

l'intera curva caratteristica. Questa è una semplice dimostrazione esemplificativa di come, in via di principio, separando le stime di uno stesso campione sulla base di diversi livelli di proprietà dei soggetti, si ottengano identiche stime, cioè i parametri dell'*item* sono invarianti rispetto al livello di abilità[82]. Infatti, tutti e quattro i procedimenti di stima, che cercano la curva con la massima verosimiglianza per ogni sottogruppo, riguardano uno stesso *item* e quindi una medesima funzione caratteristica. Ogni stima, in questo caso, parte da una sotto-sezione dell'intervallo infinito dell'abilità. Preso insieme con gli altri, ogni segmento rappresentato nei quattro grafici ricompone la curva caratteristica complessiva. La natura stessa del modello è di stimare una curva lungo il continuo infinito della proprietà, sulla base della probabilità di superare un *item* dato un certo livello di abilità: tale carattere latente non cambierà, quindi, se cambierà la posizione dei soggetti poiché "i valori dei parametri dell'*item* sono una proprietà dell'*item*, non del gruppo che risponde a esso" [Baker 2001, 55][83].

Altri argomenti possono andare a completare il quadro per giustificare e chiarire concettualmente la proprietà d'invarianza come caratteristica del modello. Una dimostrazione molto illuminante, di tipo matematico, è quella discussa da Hambleton, Swaminathan e Rogers [1991]. Applicando le proprietà dei logaritmi alla formula della probabilità, si è visto che è possibile costruire il *logit* come logaritmo del rapporto tra la probabilità di superare un *item* e quella di non superarlo (equazione 7.17). Quel risultato può essere interpretato ulteriormente come segue:

$$l\,n\frac{P(X=1)}{P(X=0)} = \beta - \delta = \alpha\beta + c \qquad \text{(eq. 7.35)}$$

dove $\alpha = 1$

$c = -\delta$.

[82] Il caso in cui lo stesso *item* ha una difficoltà diversa per individui con la stessa posizione sulla dimensione latente, ma appartenenti a gruppi diversi rispetto a certe variabili come sesso, etnia, età ecc. è noto come *differential item functioning*, presentato in modo approfondito in Tessitore [2011]. Esso non rientra, tuttavia, nella discussione attuale.

[83] Baker prosegue ricordando che "sotto la teoria classica dei test, è valido esattamente il contrario", cioè la difficoltà è "la proporzione generale di risposte corrette a un *item* per gruppo di esaminandi", per cui può variare *costitutivamente* al variare delle caratteristiche di quell'ultimo.

In quest'ultima formula s'interpreta la 7.17 come *funzione di regressione* che mette in relazione il logaritmo dell'*odd ratio* con la dimensione latente *β*, vista come variabile il cui coefficiente angolare è fisso e pari a 1 e la cui intercetta, anch'essa fissa, è rappresentata dall'opposto di *δ*, che per ogni *item* è costante. Da ciò consegue che, così come nel caso della regressione il parametro di pendenza α e l'intercetta *c* sono costanti per qualsiasi sottopopolazione individuata dai valori della variabile indipendente, così nel caso della IRT e del modello di Rasch le stime dei parametri dei soggetti e degli *item* sono indipendenti rispettivamente dagli *item* e dal campione, poiché al variare dell'abilità dei soggetti rimangono costanti il livello di difficoltà e la pendenza della retta che descrive la relazione con il *logit*. Da questo punto di vista, i modelli della IRT possono essere visti come una forma non lineare dei modelli di regressione [Hambleton 1994].

7.2.5 La logica dell'invarianza nella stima dei soggetti

Il ragionamento appena svolto vale anche, specularmente, per la valutazione dei soggetti rispetto agli *item*, il che permette di sviluppare la riflessione sul concetto di *oggettività specifica*: le caratteristiche del modello rendono possibile confrontare i soggetti in modo generalizzabile, indipendentemente dalle condizioni osservative, ossia dallo strumento di rilevazione (gli specifici *item* utilizzati).

La conseguenza diretta è chiara. Utilizzando i *logit* come unità di confronto nel modello di Rasch, le persone sono potenzialmente misurate su una scala a intervalli: si possiede un'unità lineare di suddivisione del *continuum* ed è stabilito uno zero arbitrario cui è ancorata la scala stessa. Date queste condizioni, le *differenze tra i punteggi*, a prescindere dal livello di proprietà espresso, hanno un significato invariante rispetto all'*item* utilizzato. Questo accade perché, facendo la differenza tra i *logit* delle abilità di due soggetti, sparisce il parametro dell'*item* e rimane solo la differenza pura tra i due parametri *β* sul tratto latente.

La dimostrazione è semplice e immediata. Si diano due soggetti, *p* e *q*, e un *item j*. I rispettivi *logit* saranno determinati come segue:

$$ln\left[\frac{P(X_{pj}=1}{P(X_{pj}=0}\right] = \beta_p - \delta_j \qquad \text{(eq. 7.36)}$$

$$ln\left[\frac{P(X_{qj}=1}{P(X_{qj}=0}\right] = \beta_q - \delta_j \qquad \text{(eq. 7.37)}$$

Si proceda adesso alla differenza dei due *logit*:

$$ln\left[\frac{P(X_{pj}=1}{P(X_{pj}=0}\right] - ln\left[\frac{P(X_{qj}=1}{P(X_{qj}=0}\right] = \beta_p - \delta_j - (\beta_q - \delta_j) = \qquad \text{(eq. 7.38)}$$

$$\beta_p - \delta_j - \beta_q + \delta_j = \beta_p - \beta_q \qquad \text{(eq. 7.39)}$$

La 7.39 mostra che la differenza sulla scala dei *logit* si riduce effettivamente alla differenza tra le rispettive abilità dei soggetti, senza dover dare conto dell'*item*. Accade, in via di principio, quella separazione tra misura e strumento invocata da Wright che darebbe vita a "una quantità, senza riserve per quanto riguarda quale fosse lo strumento particolare o quale fosse la particolare situazione", quella distinzione tra "occasione di misurazione" e "misura" che è appunto il cuore del concetto di *oggettività*.

Il legame di necessità che lega la sufficienza delle stime nel modello di Rasch al concetto di oggettività specifica è ricostruito da Fischer [1995a], il quale in una prospettiva assiomatica espone l'insieme di assunti da cui è possibile ricavare una funzione generale per la famiglia dei modelli di Rasch[84]. Egli rileva che è possibile arrivare a quello stesso risultato attraverso due *set* di assiomi. I primi tre sono comuni a entrambi gli insiemi[85]:

i) funzione del tratto latente continua nello spazio dei numeri reali e strettamente monotona (la IRF);
ii) codominio della funzione (la probabilità di superare l'*item*) che varia tra 0 e 1 al tendere del dominio (il tratto latente) rispettivamente a -∞ e a +∞;
iii) indipendenza locale.

[84] Si parla di "famiglia" perché, rispetto a tali assunti, il modello di Rasch rappresenta un caso particolare.

[85] Non riteniamo necessario in questo specifico contesto utilizzare la notazione tecnica completa adoperata da Fischer, per cui l'illustrazione sarà il più possibile letterale.

Se i primi tre sono considerati "assunti tecnici", il quarto è "l'assioma fondamentale e sostanziale dal quale può essere derivato il modello di Rasch" [ibidem, 461][86]:

iv) sufficienza del punteggio grezzo per la stima del tratto latente.

Non ripeteremo qui, ovviamente, i motivi per cui tale assioma costituisca il cuore del modello di Rasch. Ora, però, tenendo fermi gli assiomi i), ii) e iii), è possibile sostituire il iv) con uno che si dimostra del tutto equivalente ai fini della derivazione del modello:

v) oggettività specifica, combinata con il principio di verosimiglianza.

Fischer così definisce l'oggettività specifica[87]:

- ogni soggetto *s* è caratterizzato da un parametro di abilità β_s;
- ogni item *i* è caratterizzato da un parametro di difficoltà δ_i;
- la variabile di risposta bernoulliana (dicotomica) è caratterizzata da un parametro di risposta p_{si} determinato da una funzione definita come in i) e ii), $F(\beta_s, \delta_i)$;
- esiste una "funzione di comparazione" *U* che confronta qualsiasi coppia di soggetti, *p* e *q*, sulla base delle rispettive probabilità di risposta a un qualsiasi *item j*, tale che l'esito del confronto non dipenderà dal parametro δ_j ma solo dai parametri β_p e β_q. Tale funzione di comparazione si definisce nella forma dell'equazione 7.40:

$$U(F(\beta_p, \delta_j), F(\beta_q, \delta_j)) = V(\beta_p, \beta_q) \qquad \text{(eq. 7.40)}$$

dove U e V sono funzioni continue nello spazio dei numeri reali. Ora, osservando la struttura della 7.40, si potrà notare come essa descriva quello che accade nelle due equazioni 7.38 e 7.39, dove la scala *logit* funge da funzione di comparazione che elimi-

[86] Fischer rimanda a Rasch [1961], Birnbaum [1968], Andersen [1973], Fischer [1974].
[87] A tal proposito, si richiama esplicitamente a Rasch [1967, 1968, 1972, 1977].

na l'intervento del parametro dell'*item* dal calcolo della differenza delle abilità dei soggetti, rendendola una quantità indipendente dallo strumento[88].

Le basi strutturali e matematiche su cui poggia il modello di Rasch creano, quindi, le condizioni per generalizzare i risultati dei confronti tra i soggetti, avvicinandosi all'ideale di una misura come quantità pura e astratta dalle condizioni specifiche del processo[89]. Quanto affermato va ovviamente interpretato alla luce di quanto già discusso rispetto al particolare tipo di procedura di "misurazione", che distingue questo modello dalla forma classica di tipo concatenato.

La potenza delle proprietà del modello di Rasch non deve però essere impropriamente estesa. Nel testo abbiamo più volte tenuto a porre l'accento che i caratteri analizzati appartengono al modello in sé e ai parametri propri delle curve con cui s'ipotizzano siano in relazione il tratto latente negli *item* e nella popolazione: la nostra disamina si è concentrata su queste caratteristiche formali. Tali parametri sono ovviamente ignoti. Non a caso, il compito dell'indagine empirica è proprio quello di stimarne l'entità sulla base degli assunti del modello. I dati empirici raccolti e il modello astratto non sono necessariamente due specchi che si riflettono: la "realtà" è prima concettualizzata dal ricercatore, poi rilevata sulla base di una definizione operativa che struttura e seleziona il dato empirico e, solo alla fine, quest'ultimo può essere confrontato con il modello teorico, alla ricerca del grado di adattamento agli assunti formali e quindi del grado in cui certi caratteri possano essere estesi al dato empirico. Insomma, la stima dei nostri ormai ben noti parametri si può immaginare come una descrizione più o meno precisa (in pratica mai perfettamente esatta) della dimensione denotativa dei concetti con cui andiamo a interpretare la realtà. Il modello sarà più o meno adattato ai dati, il che significa anche che le previsioni del modello stimato avranno sempre un margine di residualità rispetto ai dati empirici. Insomma, la nostra "fotografia della realtà" (se ci è concesso usare questa metafora epistemologicamente molto ingenua), per quanto possa essere messa a fuo-

[88] Il secondo punto dell'assioma v), il principio di verosimiglianza, richiede che "*U* dovrebbe essere una funzione di una non banale (condizionale o incondizionale) verosimiglianza delle risposte" di *p* e *q* a *j*.

[89] A questo proposito, si parla di vere e proprie banche degli *item* in cui catalogare affermazioni già calibrate e utilizzabili in modo confrontabile nei più diversi contesti di ricerca. [Wright e Stone 1999; Baker 2001]

co nitidamente, non eliminerà mai del tutto una certa "sfocatura" rispetto all'immagine "vera" (sempre che ne esista una). Oppure, in altro senso, possiamo anche dire che la rigidità dei modelli astratti formali non riuscirà mai a risolvere ed esaurire completamente la variabilità di una realtà (soprattutto quella sociale) che non necessariamente rispetta in tutto e per tutto, anche solo localmente, la logica di assunti e assiomi.

Ecco perché la dimostrazione logico-matematica dell'invarianza dei parametri, della sufficienza dei punteggi, dell'oggettività specifica, non sono generalizzabili automaticamente ai dati analizzati sulla base di quel modello. Analizzeremo adesso, in conclusione, alcuni snodi critici fondamentali[90].

7.3 Metrica, oggettività specifica e rapporto tra *item* e soggetti

Fino a questo punto abbiamo lasciato sullo sfondo la questione della *metrica*, come se nel calcolo delle scale *logit* ve ne fosse una "naturale". In realtà, se nel modello di Rasch si può assumere che l'unità della scala sia pari a 1, in virtù del fatto che il coefficiente di discriminazione è costante e pari a 1, la posizione della scala è lasciata alla decisione del ricercatore: in altre parole, poiché quella riprodotta ha i caratteri di una scala a intervalli, la posizione del punto '0' e quindi l'ancoraggio della metrica è arbitraria e scaturisce dal processo di stima. Non solo, quindi, la tecnica utilizzata può avere un peso, ma hanno rilevanza soprattutto gli orientamenti teorici del ricercatore.

Infatti, in generale è possibile decidere di fissare il punto '0' ancorandolo a un gruppo di *item* o a un gruppo di soggetti. Vi possono essere ragioni teoriche molto fondate per stabilire che il punto di partenza della scala debba corrispondere a un certo livello di proprietà, espresso da alcuni *item* o da un certo gruppo di soggetti. E soprattutto possono esserci ragioni d'interpretazione, ovvero di esplicitazione del significato della scala attraverso il confronto con criteri esterni d'intelligibilità. Da questo punto di vista, riflettendo sugli elementi sempre vivi della discussione che impegnò di Whitely e Dawis e

[90] La riflessione dei seguenti paragrafi, in generale, parte soprattutto dalle riflessioni fondamentali svolte nella seconda metà degli anni '70 in un importante dibattito: Whitely e Dawis [1974, 1976], Whitely [1977], Wright [1977b]. Si vedano anche Baker [2001], de Ayala [2009] e, in italiano, Giampaglia [1990, 2008], Maggino [2004a; 2004b], Tessitore [2011].

Wright (vedi nota 90), ci pare degna di attenzione la distinzione tra posizioni diverse che tengono insieme aspetti epistemologici, teorici e tecnici.

7.3.1 Metrica e significato della misura

E' Whitely [1977] a renderle esplicite, chiamando in causa aspetti già toccati nel primo capitolo di questa tesi. In particolare, l'autrice utilizza la distinzione di Torgerson [91] tra una misurazione di tipo *fondamentale* (dove i numeri rappresentano un'operativizzazione diretta della proprietà, senza bisogno della mediazione di misure di altre proprietà), una di tipo *derivato* (dove la misurazione è mediata dalla misurazione di altre proprietà) e una definita *by fiat*[92] (dove la misurazione riguarda rapporti solo presunti tra quanto osservato e la proprietà d'interesse).

Orbene, su questa base si possono individuare (almeno) due orientamenti che si concretano in due approcci diversi ai processi di calibrazione della scala. Il primo punto di vista, più vicino all'ispirazione di uno studioso come Wright, tende a preferire gli aspetti del modello di Rasch che lo legano a una concezione "fondamentale" della misurazione. Se ciò che si sta misurando e calibrando è direttamente la proprietà latente, allora il *focus* si può concentrare sulle proprietà formali del modello, che possiamo riassumere nel concetto basilare di oggettività specifica: diventa centrale testare l'invarianza dei parametri lungo la dimensione latente e il rispetto degli assunti di base. In altre parole, si tratta sostanzialmente di verificare l'aderenza del modello ai dati. Qui è il processo di stima che fornisce gli elementi per l'interpretazione dei punteggi e la calibrazione diventa la "banale questione di trovare un punto di riferimento" [Wright 1977b]. Insomma, il procedimento tecnico della stima dei parametri e la conferma delle proprietà e degli assunti del modello sono, in un'ottica simile a quella operazionalista, i vettori del significato della scala. Il controllo dell'adattamento del modello avviene principalmente attraverso tre strategie, basate rispettivamente i) sul controllo

[91] A sua volta debitore del già richiamato contributo del grande scienziato ed epistemologo Campbell.

[92] Potremmo tradurre quest'ultima come misurazione "arbitraria", "per decreto". Torgerson [1958] ritiene di non associare il secondo tipo di misurazione (derivata) a nessun campo delle scienze sociali, poiché non esistono nessi tra asserti teorici abbastanza robusti; la misurazione fondamentale riguarderebbe invece alcune procedure di stampo sperimentale in campo psicofisico; la misurazione *by fiat* sarebbe infine quella coinvolta nei test mentali in generale, poiché il processo non avviene in ambito sperimentale e la proprietà d'interesse non è direttamente calibrata.

dell'invarianza dei parametri in vari sottogruppi di soggetti[93], ii) sull'utilizzo dei residui tra dati rilevati e previsioni del modello attraverso la costruzione di statistiche basate sul Chi-Quadrato, iii) sull'analisi dei profili individuali dei soggetti alla ricerca di vettori-persona anomali con riferimento al rapporto tra l'abilità del soggetto e le specifiche risposte agli *item* [Tessitore 2011]. Anche gli assunti di base sono sottoposti a controllo, in particolare quello di unidimensionalità[94], pur se qui aumenta esponenzialmente la varietà di procedure possibili: Maggino [2004b] riporta che sono state elencate decine di modi per avvicinare lo scopo. Si può comunque dire che, trattandosi di un'analisi di dimensionalità, molte strategie sono basate per lo più sugli strumenti e i principi dell'analisi fattoriale e, più in generale, sulle matrici di covarianza e correlazione.

Sull'altro versante troviamo la posizione più vicina alle convinzioni di studiosi come Whitely. Per chi ritiene che i modelli di cui stiamo trattando non possano essere associati al concetto di misurazione fondamentale[95], ma piuttosto a quello di misurazione *by fiat*, il ruolo della stima è altresì importante, ma non più sufficiente. Se, infatti, i rapporti tra ciò che si osserva e la proprietà che s'intende misurare non sono teoricamente abbastanza forti da andare oltre una sfera ipotetica e putativa, allora è necessario che nel processo di misurazione intervengano criteri esterni di controllo: l'oggettività specifica come *risultato tecnico* non basta più, si rivela una base troppo ristretta e si deve quindi volgere l'attenzione al concetto più ampio di *validità* e al rapporto di senso che il processo di misurazione intrattiene con costrutti esterni a esso. L'aderenza dei dati al modello conferma le proprietà e gli assunti formali che lo caratterizzano, ma non dice niente di definitivo sul fatto che ciò che si è "misurato" e la scala che si è costruita siano effettivamente la rappresentazione del costrutto latente oggetto di ricerca. Questi presupposti sfociano nella convinzione di Whitely e Dawis [1974] che l'ancoramento della scala non abbia proprio niente della banale operazione di spostamento dello zero arbitrario, bensì rappresenti la "chiave dell'interpretabilità campione-invariante dei punteggi di abilità". Per questo, in conformità a considerazioni teoriche, è compito del ri-

[93] Ma è fondamentale anche la procedura inversa, ossia il controllo dell'invarianza della stima dei parametri di abilità attraverso sottogruppi di *item*, separati con modalità randomizzate o sulla base di loro caratteristiche intrinseche.

[94] Che, ricordiamo, ha una stretta connessione con quello d'indipendenza locale.

[95] E del resto, nella nostra discussione precedente pensiamo di aver esposto almeno alcuni dei motivi per cui il modello di Rasch non può effettivamente essere accomunato alla visione classica di una misurazione fondamentale di tipo concatenato; nella Parte Terza ci occuperemo di altri modi per accertare l'additività degli attributi intensivi.

cercatore ancorare i parametri di abilità a un gruppo di soggetti o di *item* di riferimento, in virtù di una concezione più ampia di oggettività che riconosca che "l'interpretabilità dei punteggi dipende dall'avere punteggi che i ricercatori possano interpretare, per gli esaminati, come riferiti a qualità esterne al processo di misurazione" [Whitely 1977, 233][96].

7.3.2 Il presupposto operativo della generalizzabilità: l'equalizzazione dei test

La problematica della metrica è importantissima perché i principi d'invarianza e oggettività specifica (ma anche la costruzione dei *logit* stessi) partono dal presupposto che i parametri stimati giacciano su una medesima scala, cioè siano calibrati in modo del tutto confrontabile. Diversamente, anche se le stime rispecchiassero in pieno le proprietà d'invarianza, ci ritroveremmo a commentare valori diversi dello stesso parametro in situazioni diverse. Una piena generalizzabilità di stime che siano indipendenti dalla situazione di misurazione, richiede la confrontabilità delle metriche e la possibilità di trasformarle tutte in una comune. Tale procedimento di allineamento è noto come *test equating* [Lord 1980; Hambleton, Swaminathan e Rogers 1991; Baker 2001] e ha bisogno di riferimenti (che siano soggetti o che siano *item*) attorno ai quali stimare le differenze tra le calibrazioni di due o più scale a intervalli: una volta stabilito il *gap*, si possono traslare le stime in modo che coincidano su una stessa scala.

Per esempio, uno stesso campione di soggetti può avere eseguito due test, di difficoltà media diversa. Ogni test può essere stato calibrato, per esempio, in modo che la media delle difficoltà degli *item* sia pari a zero. Adesso, però, si tratta di fare in modo che l'abilità media sia uguale nei due test, perché essa è riferita allo stesso campione e deve, in linea di principio, coincidere al di là dell'occasione di misurazione. Sulla prima scala i soggetti hanno, poniamo, un'abilità media pari a 0,237, mentre sulla seconda scala gli stessi individui hanno abilità media pari a -0,089. A questo punto si deve decidere, in modo chiaramente arbitrario ma non senza ricorrere a considerazioni teoriche, quale scala sarà quella la cui metrica diverrà comune. Poniamo, quindi, che in base a riflessioni teoriche, si decida di utilizzare la metrica del secondo gruppo: è necessario a que-

[96] Notiamo di passaggio che l'autrice spinge oltre la sua critica, sostenendo che una volta assegnate le proprietà salienti del modello di Rasch al campo della misurazione fondamentale, nell'ambito della misurazione *by fiat* i test classici conserverebbero addirittura dei vantaggi e si mostrerebbero superiori per quanto riguarda l'interpretazione dei punteggi.

sto punto traslare le stime del primo gruppo di una quantità pari alla differenza tra le loro abilità medie. In questo caso, tale differenza è pari a

$$\overline{\beta_1} - \overline{\beta_2} = 0{,}237 - (-0{,}089) = 0{,}326$$

Si deve quindi trasportare ogni parametro, individuale e di *item*, di una costante pari a 0,326 per arrivare alla nuova metrica comune.

Teoricamente, calibrando ed equalizzando, si può arrivare a costruire vere e proprie banche dati di *item* pre-calibrati, pronti per essere utilizzati su campioni diversi (vedi nota 89). Di per sé, questa possibilità apre l'orizzonte a benèfici effetti nel campo della generalizzazione dei risultati e della comparazione tra campioni e popolazioni diverse. Il "sogno" di molte grandi ricerche nella storia della sociologia è stato proprio quello di poter confrontare grandi moli di dati riguardanti indagini comparative, oppure poter utilizzare strumenti comparati di testata affidabilità[97]. Qua però torna in campo, di nuovo, la grande questione dei differenziali di significato e del concetto di validità.

7.3.3 Contesto e significato in rapporto agli assunti

Uno degli assunti fondamentali del modello di Rasch, comune a quasi tutti gli approcci di misurazione, è che la risposta a un *item* sia indipendente dalla situazione, ivi comprese quindi le risposte agli altri *item*. Qui s'incontrano evidentemente dei nodi critici, riconducibili almeno a due grandi ordini di problemi.

Il primo ordine di problemi riguarda il contesto stesso della rilevazione. Proprio Whitely e Dawis [1976] sollevano, in base a ricerche *ad hoc*, il problema per il modello di Rasch dei possibili effetti dell'ordine di presentazione degli *item* e della loro sistemazione in batterie con specifiche caratteristiche, sui risultati di un test. Tale distorsione può essere ben più pericolosa nell'ambito dei test di abilità, sotto forma di effetto di apprendimento; ciò non toglie che anche nel campo di studio degli atteggiamenti le regole di somministrazione dei test possano avere ricadute sull'impatto cognitivo ed emotivo di talune affermazioni e quindi sulla risposta dei soggetti. Alla presenza di tali fenomeni, l'indipendenza locale del modello si trasforma in una chimera, così come

[97] E' questo il vasto campo delle forme equivalenti dei test mentali.

l'invarianza dei parametri stimati: principalmente perché la distorsione della difficoltà degli *item* produce effetti a catena che partono dalla calibrazione delle stime e toccano tutti gli aspetti che ne derivano più o meno direttamente.

Il secondo ordine di problemi riguarda l'inevitabile variabilità legata alla sfera dei significati. Gli *item* sono pur sempre frasi, affermazioni, definizioni, espressioni verbali veicolate con un certo linguaggio e in un determinato contesto. E quando si ha a che fare con "portatori" di significato, il ruolo di quest'ultimo non può essere in alcun modo accantonato. In particolare, in letteratura si rileva l'evenienza che individui con determinate caratteristiche (e quindi certi gruppi di individui associati a tali caratteristiche) con una stessa quantità di proprietà sul tratto latente, possano avere probabilità diverse di superare un *item*. Si fa in generale riferimento a questo fenomeno con l'espressione *differential item functioning* (già citato in nota 82). In pratica, sulla base di determinati caratteri culturali, etnici, di genere, d'istruzione ecc. la curva di probabilità di uno stesso *item* può cambiare tra diversi sottogruppi, negando quindi il principio d'invarianza del parametro δ rispetto a diversi campioni. Tale meccanismo ci pare possa presentarsi con una certa verosimiglianza nella sfera degli studi sugli atteggiamenti: la dimensione culturale ha un peso rilevante nello stabilire gerarchie di comportamento e opinione ed è acquisizione comune che in ogni sottogruppo della popolazione possano esistere riferimenti e ordinamenti di valore parzialmente diversi. Possedere lo stesso grado di tolleranza in ambienti sociali e geografici diversi, può condurre a esiti differenti nel giudicarne il livello latente in una certa affermazione a essa riferita. Per questo ogni indagine dovrebbe preoccuparsi di osservare l'eventuale incidenza di questi meccanismi, che alterano in modo radicale le pretese di oggettività e invarianza del modello di Rasch.

7.3.4 Commento finale

Queste note conclusive possono essere riassunte in tre riflessioni finali. Innanzitutto esiste un livello logico-formale del modello di Rasch, che possiede determinate caratteristiche funzionali e matematiche di evidente potenza e desiderabilità, nel solco di un ideale modello di misurazione generalizzabile e invariante. Lo studio strutturale del modello astratto serve quindi ad approfondirne le peculiarità e comprendere il significato logico dei suoi meccanismi di funzionamento.

Un secondo livello riguarda il rapporto e il legame di significato tra le operazioni ammissibili sul modello, i suoi assunti, le sue proprietà da una parte e dall'altra le caratteristiche della base empirica cui di volta in volta ci si riferisce. La questione della calibrazione e dell'ancoramento (con annesse tutte le procedure di stima dell'adattamento ecc.) svela questo rapporto e mostra i diversi approcci con cui esso è affrontato. Ci sembra pertinente guardare a tali orientamenti in una più generale ottica epistemologica, risalendo fino alle concezioni implicite o esplicite del tipo di misurazione in atto. Anche questo aspetto può ricollegarsi a quello più generale della distinzione tra l'approccio delle "strutture che generano funzioni" e quindi si riflettono in esse e quello delle "funzioni che generano strutture" e le definiscono.

Infine un terzo livello riguarda più strettamente il funzionamento dei processi cognitivi e di rilevazione dell'oggetto di ricerca. Il modello non sempre riesce a riprodurre ed esaurire ciò che si rileva della realtà empirica, né a mostrare che i suoi assunti e le sue proprietà possano essere avvicinate dalle condizioni d'indagine. La difficile "indipendenza" tra soggetti e *item*, nella forma dell'interazione nel contesto della somministrazione del test (forma interna) e dell'interazione con aspetti semantici legati all'ambiente culturale e umano più generale dei soggetti, serve proprio come esempio molto forte di come certi caratteri costitutivi della realtà possano difficilmente piegarsi a quelli dei modelli, se non con un grande sforzo d'idealizzazione.

PARTE TERZA

Il modello di Rasch e la Teoria della Misurazione Additiva Congiunta

8 Modello di Rasch e misurazione fondamentale

Nella Parte Seconda abbiamo introdotto i concetti generali della Teoria del Tratto Latente (TTL) e dei modelli probabilistici, a essa connessi, della *Item Response Theory* (IRT). In particolare, abbiamo analizzato in maggiore dettaglio le caratteristiche formali del modello a un parametro di Rasch. Esso possiede alcune peculiarità matematiche e strutturali specifiche che ne fanno, per alcuni, un esempio stocastico di quella che è stata introdotta in letteratura come "misurazione di tipo additivo congiunto" [Luce e Tukey 1964], in questo caso di un tratto latente, e fondata sull'osservazione dell'effetto combinato non interattivo della quantità di proprietà nei soggetti e negli *item* di un test. In particolare, la presenza di statistiche sufficienti e la proprietà d'invarianza dei parametri si qualificano come caratteristiche veramente desiderabili per lo sviluppo di procedure standardizzate di "misurazione" nell'ambito dei modelli stocastici.

Nell'ambito della TTL i modelli sono sviluppati in conformità a una serie di assunti generali, che si tengono assieme nel quadro di un approccio più ampio [Zand Scholten 2011]. Essi riguardano, da un lato, le risposte date da un gruppo di soggetti a un insieme di affermazioni e, dall'altro, la proprietà latente che per ipotesi dovrebbe influenzarle. Riassumendo quanto discusso nella Parte Seconda, in primo luogo il tratto latente è assunto come i) unidimensionale e ii) continuo. In secondo luogo, si assume che le risposte iii) siano associate a una distribuzione di probabilità e che iv) siano localmente indipendenti, se condizionate ai valori del tratto latente. In terzo luogo si assume che v) la relazione tra la probabilità di ogni risposta e il tratto latente sia descritta da una specifica funzione matematica e che quest'ultima vi) sia monotonica rispetto al tratto latente [ibidem].

Affinché le proprietà formali di un modello matematico, nel nostro caso quello di Rasch, possano essere attribuite in modo ragionevole e significativo anche alla struttura dei dati, è condizione necessaria che il modello si adatti loro in modo (più o meno) perfetto: al diminuire del *fit*, diventa meno plausibile rappresentare la struttura empirica su quella formale della funzione del modello. In generale, i modelli della IRT sono *data-oriented*, il che significa che i parametri sono stimati in modo tale da avvicinarsi il più possibile alla descrizione dei dati raccolti, siano essi coerenti o meno rispetto alle ipote-

si avanzate dai modelli stessi. A differenza di altri, il modello di Rasch adotta assunti più rigidi, come la costanza del parametro di discriminazione che costringe le curve caratteristiche degli *item* a disporsi parallelamente tra loro, permettendo un'interpretazione cumulativa e additiva dei parametri (abilità e difficoltà).

A questo riguardo, adottando una prospettiva consistente rispetto ai principi della TTL (come fanno ad esempio Borsboom e Mellenbergh [2004]), analizzare il livello di adattamento del modello ai dati significa testare complessivamente sul materiale empirico le ipotesi e gli assunti che vanno a costituire il modello stesso e le operazioni di ricerca nel loro insieme. Uno scarso livello nel *fit* dei dati può quindi mettere in discussione uno, alcuni o tutti gli assunti e le ipotesi, ma non è sempre possibile stabilire con certezza quali, poiché nessuno di essi è testato (e testabile) isolatamente, secondo la nota tesi di Quine-Duhem[98] [ibidem]. Il modello sarebbe quindi un "pacchetto" d'ipotesi avanzate per spiegare un fenomeno, controllando successivamente la sua applicabilità alla realtà.

8.1 La natura quantitativa delle proprietà come ipotesi empirica

Tra le ipotesi e gli assunti (esplicitati o meno), necessario a giustificare un livello di misurazione su una scala a intervalli è quello della natura quantitativa degli attributi psicologici. Lo studio approfondito dei caratteri matematici formali di un modello come quello di Rasch e delle trasformazioni ammissibili sulle scale, non ci consente di affrontare fino in fondo e risolvere la questione epistemologica essenziale, cioè se staremmo compiendo operazioni di rilevazione su una struttura empirica quantitativa e continua

[98] Secondo Duhem [1914, 187, cit. in Michell 2004] lo scienziato "non può mai sottoporre un'ipotesi isolata alla prova sperimentale, ma solamente un intero gruppo d'ipotesi; quando l'esperimento è in disaccordo con le sue previsioni, ciò che egli apprende è che almeno una delle ipotesi che costituisce questo gruppo è inaccettabile e dovrebbe essere modificata; ma l'esperimento non designa quale dovrebbe essere cambiata." Michell [2004, 123] è però in disaccordo con l'interpretazione di Borsboom e Mellenbergh, sottolineando che "Duhem pensava che le ipotesi non potessero essere provate isolatamente, nel senso che testare richiede sempre di congiungere le ipotesi con altre proposizioni per dedurre le previsioni, ma egli non intendeva che le ipotesi non potessero essere provate isolatamente nel senso che una prova sperimentale non possa mai essere di una sola ipotesi." Insomma, se si è sicuri della validità delle proposizioni congiunte, allora un esperimento può provare la validità o meno anche di una singola ipotesi empirica.

rappresentabile lungo i numeri reali, oppure se si tratti solo di una struttura al più ordinale.

Sulla pretesa di quantificazione dei modelli psicometrici e sulla netta distinzione logica ed empirica tra strutture quantitative e strutture ordinali (con le seconde che non implicano necessariamente le prime), si concentra la critica di alcuni autori, come si è già in parte discusso nel primo capitolo. Uno degli attacchi più forti è quello articolato in più occasioni proprio da Michell [es. 2000; 2008a; 2009], il quale denuncia quella che ritiene essere la "condizione patologica" delle scienze psicometriche. Perché una scienza possa definirsi patologica[99], secondo Michell, è necessario che nel suo "flusso normale" occorrano delle interruzioni (*breakdowns*). Può accadere, infatti, che si formino e siano portate avanti delle ipotesi false o non testate empiricamente che, a differenza di ciò che la pratica scientifica continuamente richiederebbe, non sono adeguatamente controllate, individuate ed eventualmente accertate, rimosse o corrette. Se la presenza d'ipotesi non accertate è il primo presupposto di una scienza patologica, essa richiede un secondo fattore: il campo conoscitivo in questione deve disporsi in modo tale da rendere *impossibile riconoscere* le ipotesi in questione e quindi se necessario *rimuoverle*, per tornare a un normale svolgimento dell'attività scientifica. La patologia non risiede, quindi, tanto nell'esistenza di un'ipotesi non controllata e potenzialmente fallace, ma soprattutto dall'azzeramento in seno a un settore scientifico degli "anticorpi" che dovrebbero essere in grado di individuare, affrontare e superare quel possibile errore.

A finire sul banco degli imputati, nel campo degli studi sulle dimensioni attitudinali, e a portare Michell a definirlo "patologico" è proprio l'ipotesi che le proprietà indagate sarebbero quantitative. Scrive Michell [2000, 650] che nelle scienze psicometriche

> (a) un'ipotesi basilare, empirica (cioè a dire l'ipotesi che gli attributi psicologici sono quantitativi) è accettata come vera senza che ne sia

[99] Ci risulta che l'etichetta di "scienza patologica" sia stata coniata in precedenza, nel 1953, dal chimico Irving Langmuir in un'accezione in parte diversa da quella di Michell e riferita alle scienze fisiche e naturali. Con essa s'intendeva la persistenza, in alcuni membri di una comunità scientifica, di convinzioni, idee e obiettivi che fossero *già stati* dimostrati fallaci o impossibili e trattati come tali dalla maggioranza degli altri componenti. Si tratta di un processo psicologico per cui si accettano formalmente i presupposti e le regole del metodo scientifico, salvo poi violarli inconsapevolmente con pretese che non possono essere scientifiche. In questo concetto ci sono quindi alcune connessioni con quelli celeberrimi di "paradigma" e "scienza normale" dello storico della scienza Kuhn [1970].

> mai stata testata seriamente l'adeguatezza e (b) il fatto che questa ipotesi non sia mai stata testata soddisfacentemente è tenuto nascosto.

Gli studiosi degli atteggiamenti, delle opinioni, dei valori avrebbero, secondo quanto sostiene Michell [2008a], applicato la teoria della misurazione di Stevens, concentrandosi sul versante formale dei sistemi relazionali, dando per scontato di stare misurando proprietà quantitative continue su scale a intervalli, senza però adoperarsi per provare efficacemente tale assunto[100]. Michell ipotizza che i motivi di tale omissione risiederebbero nella volontà di presentare la propria disciplina come affine, nei metodi e negli esiti, alle scienze fisiche e naturali, secondo un atteggiamento etichettato come "scientismo"[101]. In questo contesto, ai modelli probabilistici della IRT si riconosce un ruolo per certi aspetti più avanzato rispetto a quelli della Teoria classica dei test, basati principalmente sui punteggi grezzi. Eppure anche nelle ricerche che utilizzano i primi, l'assunto del carattere quantitativo degli attributi è dato praticamente sempre per scontato, senza testare l'ipotesi sul campo.

Come si è avuto modo di ripetere a più riprese (e come dovrebbe essere stato confermato in maniera abbastanza chiara attraverso lo studio del modello di Rasch), il tipo d'indagine scientifica che coinvolge le dimensioni psicologiche non utilizza procedure assimilabili a quelle abbracciate dalla concezione classica della misurazione fondamentale estensiva: se anche fossero grandezze quantitative, gli attributi psichici non lo sarebbero comunque di un tipo concatenabile. Non possediamo e non siamo in grado di definire teoricamente un'unità di misura che possa essere idealmente giustapposta

[100] Michell [2007] riporta alcuni esempi di conoscenza convenzionale, attestazioni sulla natura intervallare della misurazione senza che siano avanzate giustificazioni nemmeno per una di tipo ordinale: "la misurazione intervallare è probabilmente la scala più comune in psicologia" [Lehman 1991, 54]; "gran parte delle misure di stati e tratti psicologici e dei costrutti quali gli atteggiamenti e le interpretazioni di eventi delle persone sono al livello di intervalli" [Whitely 1996, 117]; "è elevata la probabilità che molte scale e test usati nella misurazione psicologica ed educazionale approssimino la misurazione intervallare" [Kerlinger e Lee 2000, 635]; "la vasta maggioranza dei test psicologici che misurano intelligenza, abilità, personalità e motivazione… sono scale a intervalli" [Kline 2000, 18]. A questi esempi, aggiungiamo anche Lord e Novick [1968, 21]: "Il livello di misurazione specificato più spesso nella teoria dei test mentali è la misurazione a intervalli, che produce una scala a intervalli."

[101] Michell avanza anche una seconda motivazione, di carattere economico: il prestigio di una scienza e l'autorevolezza imputata ai suoi risultati sono elementi importanti anche per il reperimento di fondi e il sostegno finanziario alla ricerca accademica [cfr. anche Heene 2013]. Come si capisce, quello di Michell è un attacco portato con grande forza e su più fronti.

all'oggetto da misurare un numero *n* di volte. Il tipo di "misurazione" delle dimensioni psicologiche latenti coinvolta nel modello di Rasch, concedendo che di misurazione in senso quantitativo si tratti, è di natura diversa perché utilizza l'osservazione dell'effetto congiunto di due variabili per stimare la posizione dei soggetti sul continuo latente incognito.

Parallelamente allo sviluppo dei modelli psicometrici probabilistici, nell'ultimo mezzo secolo, a partire dall'ambito della Teoria rappresentazionale della misurazione (*Representational Measurement Theory*, RMT), sono stati compiuti sforzi teorici robusti per studiare e approfondire i presupposti necessari affinché, *anche* nelle condizioni particolari in cui si trova la ricerca sugli atteggiamenti, sia possibile dimostrare in via assiomatica che un certo tratto latente sia effettivamente quantitativo. I risultati di tale opera di elaborazione sono individuati da quella che è definita "Teoria della misurazione additiva congiunta" (in inglese *Additive Conjoint Measurement Theory*, ACM). Individuare un modello di misurazione in grado di affrontare gli assunti dell'ACM significherebbe aprire la strada alla possibilità di testare empiricamente l'ipotesi che una certa struttura empirica latente possieda i caratteri di un attributo quantitativo. La critica di Michell è che, pur avendo ora a disposizione gli strumenti per testare le ipotesi sulla natura delle proprietà psicologiche[102], gli scienziati della sfera umana abbiano continuato a "nascondere la polvere sotto il tappeto", ricercando raramente e in modo insoddisfacente conferme.

A questo proposito, una serie di studiosi (come vedremo tra poco) ha individuato proprio nel modello di Rasch un caso di misurazione additiva congiunta, in una versione probabilistica, suscitando però perplessità in altri autori. La discussione, da questo punto in poi, si svilupperà su questo nodo: può il modello di Rasch rappresentare un *trait d'union* tra approcci diversi al concetto di misurazione e può, attraverso la logica dell'ACM, aiutare a scogliere il nodo della natura delle dimensioni psicologiche? Cominciamo illustrando le caratteristiche e gli assunti della misurazione additiva congiunta.

[102] Michell [2009] tiene a sottolineare che il suo richiamo all'ACM non significa che egli accolga in generale i presupposti teorico-concettuali del *frame* rappresentazionalista, in cui gli assiomi della misurazione congiunta sono stati, come detto, sviluppati.

9 La teoria della misurazione additiva congiunta

Si è visto nel primo capitolo che una delle critiche più diffuse rivolte alla teoria delle scale di Stevens (teoria di stampo rappresentazionale e operazionalista) riguarda la centralità, per stabilire il livello di misurazione, delle operazioni ammissibili sulle scale che ne mantengano invariante la struttura delle relazioni. Potremmo riassumere la critica in questa forma: tra la funzione di rappresentazione (che connette la struttura relazionale empirica con quella formale) e quella di unicità (che stabilisce le trasformazioni ammissibili sulla scala formale), alla seconda è dato un peso maggiore che alla prima, rischiando di rendere "misurazione" qualsiasi assegnazione numerica e di mettere in secondo piano la reale consistenza della scala formale rispetto alla struttura empirica. Questo elemento si associa, nelle scienze umane, alla difficoltà di giustificare empiricamente il carattere quantitativo degli attributi mentali, che non riguardano grandezze concatenabili[103] estensive, bensì intensive.

Nonostante già Hölder all'inizio del XX secolo avesse chiarito in generale i requisiti di una struttura additiva, è dagli anni Sessanta che si presenta in forma completa una teoria della misurazione congiunta che, per via algebrica e assiomatica, dimostra in quali condizioni e secondo quali presupposti, attributi non concatenabili possono comunque essere accertati come quantificabili e additivi. A inaugurare in modo organico questa nuova pagina della Teoria rappresentazionale della misurazione (RMT) sono Luce e Tukey [1964] sul *Journal of Mathematical Psychology*, seguiti da altri contributi che hanno ampliato e approfondito l'orizzonte della teoria [Scott 1964; Krantz 1964] fino all'opera di Krantz, Luce, Suppes e Tverski [1971] sui fondamenti della misurazione, che sviluppa a largo raggio la teoria assiomatica della misurazione e fornisce giustificazioni sulla base di una raffinata e complessa rete di dimostrazioni algebriche.

Il lavoro di Luce e Tukey [1964] ebbe una portata secondo molti potenzialmente rivoluzionaria, non da ultimo perché dimostrava che anche gli attributi psicologici, intensivi e non concatenabili, possono essere legittimamente ipotizzati come quantitativi e continui e che è possibile testare empiricamente tale ipotesi attraverso la soddisfazione di una serie di assiomi. La tesi di N. R. Campbell, certificata dalla Commissione Fergu-

[103] Si ricordi la discussione avvenuta nella Commissione Ferguson (vedi cap. 1).

son poco più di venti anni prima, era in via teorica messa in seria discussione, se non smentita.

Per impostare l'illustrazione della teoria[104], cominciamo presentando la tabella 9.1. Qui troviamo due attributi, *A* e *M*, dei quali non dobbiamo conoscere necessariamente la natura ordinale o quantitativa, ma di cui sia possibile però distinguere e identificare vari livelli tra loro indipendenti, per cui *A* sarà composto dai livelli *a*, *b*, *c* ecc. e *M* dai livelli *m*, *n*, *o* ecc.

		M					
		m	*N*	*o*	…	…	…
A	*a*	*(a,m)*	*(a,n)*	*(a,o)*	…	…	…
	b	*(b,m)*	*(b,n)*	*(b,o)*	…	…	…
	c	*(c,m)*	*(c,n)*	*(c,o)*	…	…	…
	…	…	…	…	…	…	…
	…	…	…	…	…	…	…
	…	…	…	…	…	…	…

Tab. 9.1 – Tabella dei tre attributi *A*, *M* e *Z*.

Il terzo attribuito presente nella tabella, *Z*, è il risultato di una combinazione *non interattiva* di *A* e *M* nelle coppie ordinate dei loro livelli: *(a,m)*, *(b,m)*, *(a,n)* ecc. *Z* è quindi il prodotto cartesiano di *A* e *M*. I tre attributi sono considerati contemporaneamente (congiuntamente, appunto)[105].

[104] Per la ricostruzione della teoria ci baseremo principalmente, tra gli altri, oltre che sui testi già citati di Luce e Tukey [1964], Scott [1964], Krantz [1964], Krantz, Luce, Suppes e Tversky [1971], anche su Green [1986], Burro [2007; 2009], Kyngdon [2008a; 2008b], Michell [2008b; 2009], Zand Scholten [2011], Noventa e Vidotto [2012a; 2012b].

[105] In concreto, *A* e *B* possono essere costituiti da insiemi disgiunti di oggetti, i quali esprimono un certo livello dell'attributo in questione. In questo senso, nella logica della presente esposizione si può parlare di livello dell'attributo o di oggetto, elemento (con quel determinato livello di attributo) in modo pressoché indifferente.

Gli assiomi della ACM descrivono le relazioni che devono sussistere tra i livelli di *Z* affinché gli altri due attributi possano essere dimostrati quantitativi continui e additivi rispetto a *Z*. Nello specifico, è possibile dire che $Z = \langle A \times M, \geq \rangle$ è una struttura empirica di misurazione congiunta additiva se e solo se i seguenti assiomi sono verificati:

1) ordine parziale;
2) indipendenza;
3) doppia cancellazione;
4) risolvibilità;
5) condizione archimedea.

Cercheremo ora di capire il significato degli assiomi e il loro ruolo nell'accertamento del carattere quantitativo e continuo degli attributi, provando a esporre il tutto nel modo più sintetico e semplificato possibile.

9.1 Il primo assioma: l'ordine parziale

Gli attributi *A* e *M*, rispetto ai loro propri livelli, sono ordinati parzialmente se e solo se, date le coppie ordinate di *Z (a,m), (b,m)* ecc., si verifica che:

i. dati i livelli *a* e *b* in *A* e *m* in *M* e dato $(a,m) \geq (b,m)$, allora $a \geq b$;
ii. dati i livelli *m* e *n* in *M* e *a* in *A* e dato $(a,m) \geq (a,n)$, allora $m \geq n$;
iii. la relazione d'ordine '≥' è transitiva (se $a \geq b$ e $b \geq c$, allora $a \geq c$) e connessa (cioè qualsiasi livello di *A* è confrontabile con qualsiasi livello di *M* e viceversa).

In pratica, i livelli degli attributi (o gli oggetti degli insiemi *A* e *M*) formano un ordine parziale, sono cioè ordinabili secondo una relazione '≥'.

9.2 Il secondo assioma: l'indipendenza

L'assioma di indipendenza è verificato dalla relazione d'ordine '≥' sopra *A* x *M* se e solo se:

i. dati i livelli (o gli oggetti) *a* e *b* in *A* e il livello (o l'oggetto) *m* in *M*, allora *(a,m)* ≥ *(b,m)* implica che, per qualsiasi altro elemento *p* in *M*, *(a,p)* ≥ *(b,p)*;
ii. dati i livelli *m* e *n* in *M* e il livello *a* in *A*, allora *(a,m)* ≥ *(a,n)* implica che, per qualsiasi altro elemento *d* in *A*, *(d,m)* ≥ *(d,n)*.

In altre parole, l'assunto d'indipendenza stabilisce che, presi due livelli qualsiasi di un attributo, la loro relazione d'ordine non deve cambiare una volta che siano accoppiati con uno qualsiasi dei livelli del secondo attributo. L'ordinamento delle coppie, cioè, deve essere indipendente dal livello del secondo attributo e, quindi, deve rimanere costante per ogni livello di quest'ultimo.

Una conseguenza dell'assunto d'indipendenza è rilevata da Green [1986, 142], quando afferma che "una matrice dei dati è additiva solo se le sue righe e le sue colonne possono essere permutate in modo che tutti gli ingressi siano monotonicamente crescenti da sinistra a destra e dall'alto in basso, o da destra a sinistra e dal basso in alto". L'assioma d'indipendenza garantisce questa permutabilità.

Al suo riguardo si parla anche di "assioma della cancellazione singola". Infatti, il mantenimento dell'ordine tra due livelli *a* e *b* in *A*, a prescindere dal livello di *M* cui sono confrontate, è garantito dalla soddisfazione della diseguaglianza

$$(a,m) \geq (b,m) \qquad \text{(eq. 9.1)}$$

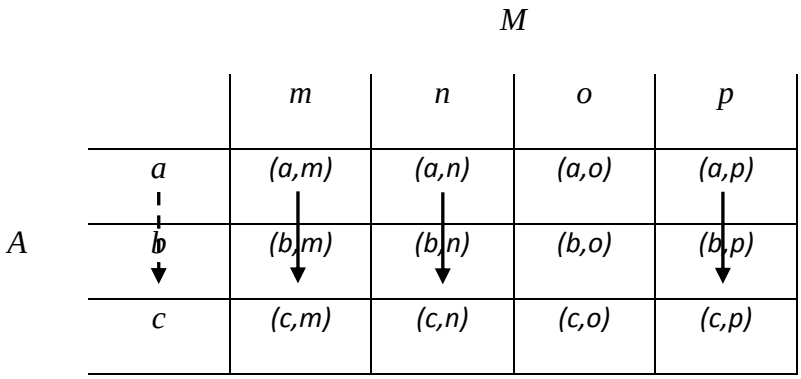
per qualsiasi livello *m* in *M*. Dato che l'eq. 9.1 è vera, in condizioni di additività, se

$$a + m \geq b + m \qquad \text{(eq. 9.2)}$$

allora, semplificando, si ottiene

$$a \geq b \qquad \text{(eq. 9.3)}$$

Si parla di cancellazione singola perché eliminando dalla diseguaglianza 9.2 il singolo fattore *m*, rimane invariata la relazione d'ordine tra *a* e *b*. La tabella 9.2 illustra l'assunto d'indipendenza sotto forma di cancellazione singola.



Tab. 9.2 – Assunto di cancellazione singola sulla porzione di una matrice. Le frecce con tratto continuo rappresentano le relazioni d'ordine antecedenti (se…), la freccia tratteggiata rappresenta la relazione d'ordine conseguente (allora).

Lo stesso procedimento vale, *mutatis mutandis*, per i livelli di *M* se confrontati con qualsiasi livello di *A*.

9.3 Il terzo assioma: la cancellazione doppia

L'assunto di cancellazione doppia è soddisfatto dalla relazione d'ordine '≥' sopra *A* x *M* se e solo se, per qualsiasi *a*, *b* e *c* in *A* e qualsiasi *m*, *n* e *o* in *M*, se $(a,n) \geq (b,o)$ e $(b,m) \geq (c,n)$, allora $(a,m) \geq (c,o)$; e se $(a,n) \geq (b,m)$ e $(b,o) \geq (c,n)$, allora $(a,o) \geq (c,m)$.

Si tratta di un assunto importantissimo, centrale nell'accertamento dell'additività del modello. Esso, infatti, sancisce sostanzialmente che se alcune relazioni d'ordine sussistono in una certa direzione, allora è necessario che anche altre relazioni d'ordine sussistano, e nella stessa direzione, affinché il modello sia una rappresentazione additiva di quantità continue misurabili su scale a intervalli.

L'assioma è definito di "cancellazione doppia" perché, partendo da due diseguaglianze, due fattori scompaiono risultando in una terza diseguaglianza. Partendo dalla prima fattispecie (che chiameremo "direzione 1"), che riguarda le relazioni d'ordine lungo le *diagonali principali* (che partono da sinistra), le due diseguaglianze

$$(a,n) \geq (b,o) \qquad \text{(eq. 9.4)}$$

$$(b,m) \geq (c,n) \qquad \text{(eq. 9.5)}$$

sono vere, ipotizzando l'additività, se

$$a + n \geq b + o \qquad \text{(eq. 9.6)}$$

$$b + m \geq c + n \qquad \text{(eq. 9.7)}$$

da cui, sommando membro a membro, si ottiene la nuova diseguaglianza

$$a + n + b + m \geq b + o + c + n \qquad \text{(eq. 9.8)}$$

che si semplifica dai livelli *b* e *n*, diventando

$$a + m \geq c + o \qquad \text{(eq. 9.9)}$$

permettendoci di concludere, come si voleva dimostrare, che

$$(a,m) \geq (c,o) \qquad \text{(eq. 9.10)}$$

E' possibile osservare, nella tabella 9.3, la logica dell'argomento, con le due frecce continue che rappresentano le relazioni d'ordine antecedenti, che implicano la terza relazione d'ordine (freccia tratteggiata).

		M		
		m	*n*	*o*
A	*a*	*(a,m)*	*(a,n)*	*(a,o)*
	b	*(b,m)*	*(b,n)*	*(b,o)*
	c	*(c,m)*	*(c,n)*	*(c,o)*

Tab. 9.3 – Assunto di cancellazione doppia sulla porzione 3x3 di una matrice. Direzione 1 (diagonali di sinistra). Le frecce con tratto continuo rappresentano le relazioni d'ordine antecedenti (se…), la freccia tratteggiata rappresenta la relazione d'ordine conseguente (allora).

Per quanto riguarda la seconda direzione (che chiameremo "direzione 2"), quella delle *diagonali secondarie* (che partono da destra), troviamo le due diseguaglianze antecedenti

$$(a,n) \geq (b,m) \qquad \text{(eq. 9.11)}$$

$$(b,o) \geq (c,n) \qquad \text{(eq. 9.12)}$$

che implicano

$$a + n \geq b + m \qquad \text{(eq. 9.13)}$$

$$b + o \geq c + n \qquad \text{(eq. 9.14)}$$

Sommando membro a membro, si ottiene

$$a + n + b + o \geq b + m + c + n \qquad \text{(eq. 9.15)}$$

e semplificando, con la scomparsa ancora di *n* e *b*, si arriva a

$$a + o \geq c + m \qquad \text{(eq. 9.16)}$$

Anche in questo caso si può quindi concludere che

$$(a,o) \geq (c,m) \qquad \text{(eq. 9.17)}$$

La tabella 9.4 mostra, nel caso delle diagonali di destra, il funzionamento dell'assunto della cancellazione doppia.

		M		
		m	*n*	*o*
A	*a*	*(a,m)*	*(a,n)*	*(a,o)*
	b	*(b,m)*	*(b,n)*	*(b,o)*
	c	*(c,m)*	*(c,n)*	*(c,o)*

Tab. 9.4 – Assunto di cancellazione doppia sulla porzione 3x3 di una matrice. Direzione 2 (diagonali di destra). Le frecce con tratto continuo rappresentano le relazioni d'ordine antecedenti (se…), la freccia tratteggiata rappresenta la relazione d'ordine conseguente (allora).

Le relazioni d'ordine prescritte dall'assunto di doppia cancellazione debbono valere per tutte le possibili sottomatrici di 3x3 elementi.

9.3.1 La cancellazione doppia nel rapporto tra ordine e quantità

Notiamo un aspetto molto importante, che ci aiuterà a fare luce sul concetto di quantità continue additive e, contemporaneamente, anche su quello della differenza strutturale tra *ordine* e *quantità* [Michell 2008a, 2008b; Kyngdon 2008b]. Come abbiamo detto, il rispetto dell'assioma d'indipendenza assicura che le righe e le colonne siano permutabili in modo tale che i livelli siano ordinati monotonicamente. Orbene, si può facilmente comprendere che in questo caso le relazioni d'ordine delle diagonali di sinistra (direzione 1, tab. 9.3) sono automaticamente verificate [Michell 1988]. Poniamo il caso di ordinare gli attributi *A* e *M* in modo decrescente, con $a \geq b \geq c$ e $m \geq n \geq o$[106]. In questo caso, spostandosi lungo le diagonali di sinistra si incontrano livelli dell'attributo *Z* che sono l'effetto della combinazione di livelli decrescenti sia di *A* sia di *M*. Risulta quindi intuitivo che sia necessario che anche i livelli di *Z*, lungo quelle direttrici, siano decrescenti, per poter effettivamente riconoscere il principio di additività. La dimostrazione matematica formale è presto fatta. Se è vero che $a \geq b$ e che $n \geq o$ allora, sommando a membro a membro, sarà del pari vero che $a + n \geq b + o$. Quest'ultima espres-

[106] Se fossero ordinati in modo crescente, basterebbe invertire il senso delle frecce oppure, che è lo stesso, interpretare l'ordine come '<' invece che '≥', per mantenere intatto il ragionamento seguente.

sione altro non è se non l'eq. 9.6. Allo stesso modo, poiché $b \geq c$ e $m \geq n$, allora è necessario che $b + m \geq c + n$, che altro non è se non l'eq. 9.7. Ciò infine vale anche per l'eq. 9.9, che si omette di mostrare per esteso perché dimostrabile esattamente come sopra. Tutte le tre diseguaglianze (le due antecedenti e la conseguente) della direzione 1 dell'assioma di cancellazione doppia seguono direttamente dall'ordinamento dei livelli.

Questo significa che, se vale l'assioma d'indipendenza e gli attributi *A* e *M* sono ordinabili indipendentemente l'uno dall'altro, allora è conseguenza necessaria che siano automaticamente verificate le relazioni d'ordine della cancellazione doppia che riguardano le diagonali di sinistra (direzione 1) delle matrici 3x3.

Questo automatismo algebrico vale anche per gli antecedenti e i conseguenti delle diagonali in direzione opposta? La risposta è no, e ora vedremo perché. Com'è immediato comprendere, ci troviamo di fronte (tab. 9.4) a una situazione nettamente diversa. Infatti, spostandoci lungo le diagonali di destra (che scendono verso il basso a sinistra) incontriamo livelli di *Z* che sono il frutto della combinazione di livelli decrescenti di *A* e di livelli crescenti di *M*. Questa associazione composita di livelli che salgono e altri che scendono non permette, già solo intuitivamente, di avere la certezza che siano rispettate le condizioni che assicurano la presenza di una struttura quantitativa continua e additiva. Infatti, se resta vero che $a \geq b$, qui troviamo che $n < m$, da cui non può direttamente discendere che $a + n \geq b + m$ (eq. 9.13). Lo stesso vale per l'eq. 9.14 ($b + o \geq c + n$), poiché $b \geq c$, ma $o < n$. Analogamente non si può concludere che $a + o \geq c + m$, perché $o < m$.

In questo secondo scenario, ci sono in via teorica casi in cui la cancellazione doppia sarà verificata e altri in cui non lo sarà. Nel primo caso i nostri attributi saranno quantitativi e continui, nel secondo caso non potremo invece andare al di là della valutazione della loro natura ordinale. Di sicuro, la semplice informazione sulle relazioni d'ordine tra i livelli degli attributi non è sufficiente all'accertamento della loro eventuale natura quantitativa, ma è necessario chiamare in causa le relazioni additive che governano le grandezze quantitative, partendo dalla basilare definizione di Euclide [Michell 2008a; si veda anche la Parte Prima del presente lavoro], per cui date le grandezze *x*, *y*, *w*, *z* ecc. e i loro reciproci rapporti, allora $x/y > w/z$ se e solo se esistono due numeri

interi, *q* e *t*, per cui valga $qx > ty$ e $qw \leq tz$, cioè $x/y > t/q \geq w/z$. E non è detto che tali numeri interi esistano.

Per usare le parole di Michell [2008a, 17]

> l'insieme di tutte le coppie ordinate di rapporti ricade con nettezza in due classi: quella in cui la relazione d'ordine tra le coppie di rapporti è determinata dall'ordine delle grandezze coinvolte (cioè, la classe 1); e quella in cui la relazione d'ordine tra le coppie di rapporti è determinata dalla struttura delle grandezze al di là del semplice ordine, ciò che potremmo chiamare la *struttura additiva dell'attributo* (cioè, la classe 2).

Pertanto, l'assioma di cancellazione doppia gioca un ruolo importante non solo nell'accertamento del carattere quantitativo continuo nell'ambito della teoria della misurazione additiva congiunta, ma più in generale nella riflessione sulle caratteristiche degli attributi in un contesto di ricerca come quello delle scienze psicologiche e sociali.

Da un punto di vista pratico, se si dimostra verificato l'assioma d'indipendenza, allora sarà sufficiente controllare le relazioni della sola direzione 2 per accertare il soddisfacimento della cancellazione doppia. Oppure, detto altrimenti, se si deciderà di operare prima un controllo integrale delle relazioni che soddisfano la doppia cancellazione, allora in caso di successo sarà verificato automaticamente anche l'assioma d'indipendenza: il primo dei due, in parole povere, sussume il secondo. Sulla base delle motivazioni riassunte in questo paragrafo, Michell [2009] dichiara che le relazioni che mostrano una struttura ordinata (direzione 1) rappresentano una metà di ciò che serve a disvelare una struttura quantitativa e che l'ordine è quantità solo a metà.

9.4 Il quarto assioma: la risolvibilità

Questo assioma stabilisce che la relazione '$\geq$' in $A \times M$ è risolvibile se, dato qualsiasi terzetto selezionato tra quattro elementi *a* e *b* in *A* e *m* e *n* in *M*, il quarto esiste sempre tale che $(a,n) \sim (b,m)$.

Secondo questo assioma, per ogni livello di *Z* deve esistere un corrispondente livello in *A* e in *M* e viceversa: non possono esistere "buchi". Una configurazione simile dei livelli suggerisce che essi debbano possedere una delle due caratteristiche seguenti [Krantz, Luce, Suppes e Tverski 1971]: o essere tra loro perfettamente equidistanti, come la serie dei numeri naturali; oppure essere densi come i numeri reali, per cui dati due qualsiasi livelli ordinati, esiste sempre un livello intermedio.

9.5 Il quinto assioma: condizione archimedea

Il quinto e ultimo assioma serve ad assicurare che nessun livello degli attributi *A* e *M* sia infinitamente piccolo o grande rispetto a qualsiasi altro, cioè che qualsiasi serie standard strettamente limitata dei livelli di *A* e *M* sia finita.

Formalmente, siano dati gli elementi *a*, *b*, *c* e *d* in *A* e gli elementi *m*, *n*, *o* e *p* in *M*. Se $b - a < d - c$ e $n - m < p - o$, allora per un qualche numero naturale *x*, *A* e *M* sono Archimedei se e solo se $x(b - a) \geq d - c$ e $x(n - m) \geq p - o$. In pratica, per quanto più piccola possa essere la differenza tra alcuni livelli rispetto a quella tra altri, sarà sempre possibile moltiplicare la prima un numero finito di volte tale da sopravanzare la seconda.

Visto da un'altra visuale, l'assioma esige che, ordinati i livelli di *A* e di *M* monotonicamente secondo la progressione dei numeri naturali, dati due livelli a_i e a_{i+1} in *A*, allora perché si verifichi che $(a_i,x) \sim (a_{i+1},y)$ è necessario che $x < y$. In altre parole, per ottenere lo stesso risultato nell'accoppiamento dei livelli, in termini di livelli indicizzati se uno aumenta di livello è necessario che l'altro diminuisca[107].

[107] Si noti la connessione tra questo assioma e quello della cancellazione doppia per la direzione 2 (un attributo aumenta, uno diminuisce e il prodotto cartesiano non deve essere inferiore) e con l'assioma di risolvibilità che richiede attributi continui e densi come i numeri reali o con livelli equidistanti e associabili ai numeri naturali.

9.6 Gerarchia delle cancellazioni

Gli ultimi due assiomi presentati (risolvibilità e condizione archimedea) non sono direttamente testabili empiricamente su insiemi composti di un numero finito di elementi (livelli). Se un attributo quantitativo continuo ha, per definizione, un numero infinito di stati possibili, da un punto di vista empirico gli stati rilevati sono necessariamente finiti[108]. Un modo alternativo di accertare gli assiomi su un numero finito di livelli dell'attributo risiede nella soddisfazione della gerarchia delle condizioni di cancellazione di ordine *s* [Scott 1964]. Definiamo *s* come il minimo tra il numero dei livelli dell'attributo *A* e dell'attributo *M*, meno uno. Definito a_i come il numero di livelli di *A* e m_j come il numero di livelli di *M*, allora

$$s = \min\{(a_i - 1), (m_j - 1)\} \qquad \text{(eq. 9.18)}$$

Se, per esempio, abbiamo una matrice 3x3, allora la gerarchia delle condizioni di cancellazione sarà di grado 3 – 1 = 2, cioè la doppia cancellazione. Se invece la matrice avesse quattro livelli per ogni attributo (4x4), allora avremmo la possibilità di testare empiricamente una gerarchia di cancellazione di grado 3, ossia una cancellazione tripla. In questo caso, la logica del controllo è illustrata graficamente dalla tabella 9.5.

Nella tripla cancellazione le tre diseguaglianze di partenza sono del tipo

$$(a,n) \geq (b,m) \qquad \text{(eq. 9.19)}$$
$$(b,o) \geq (c,n) \qquad \text{(eq. 9.20)}$$
$$(c,p) \geq (d,o) \qquad \text{(eq. 9.21)}$$

che, stante l'additività degli attributi, implicano

$$a + n \geq b + m \qquad \text{(eq. 9.22)}$$
$$b + o \geq c + n \qquad \text{(eq. 9.23)}$$
$$c + p \geq d + o \qquad \text{(eq. 9.24)}$$

[108] Perché finito è il numero di soggetti su cui si rilevano gli stati sulle proprietà attitudinali e il numero di *item*.

Sommando a membro a membro le tre diseguaglianze, si ottiene la 9.25:

$$a + n + b + o + c + p \geq b + m + c + n + d + o \qquad \text{(eq. 9.25)}$$

Nella tripla cancellazione scompaiono, per semplificazione, i livelli *b*, *c*, *n* e *o*, giungendo infine a

$$a + p \geq m + d \qquad \text{(eq. 9.26)}$$

Dalla 9.26, si deduce pertanto la condizione conseguente della cancellazione tripla, cioè

$$(a,p) \geq (d,m) \qquad \text{(eq. 9.27)}$$

		M			
		m	*n*	*o*	*P*
A	*a*	*(a,m)*	*(a,n)*	*(a,o)*	*(a,p)*
	b	*(b,m)*	*(b,n)*	*(b,o)*	*(b,p)*
	c	*(c,m)*	*(c,n)*	*(c,o)*	*(c,p)*
	d	*(d,m)*	*(d,n)*	*(d,o)*	*(d,p)*

Tab. 9.5 – Cancellazione tripla sulla porzione 4x4 di una matrice. Direzione 2 (diagonali di destra). Le frecce con tratto continuo rappresentano le relazioni d'ordine antecedenti (se…), la freccia tratteggiata rappresenta la relazione d'ordine conseguente (allora).

Via via che aumenta l'*s*-esimo ordine della gerarchia di cancellazione, si procede con la medesima logica, allargando progressivamente il numero dei possibili livelli coinvolti.

Se i cinque assiomi che abbiamo illustrato sono empiricamente verificati, allora si può concludere che gli attributi *A* e *M* possiedono una struttura (non concatenabile ma

additiva) sufficientemente complessa da poter sostenere una rappresentazione sull'insieme dei numeri reali. Infatti, si può a quel punto ammettere che gli attributi sono ordinabili, che il loro ordinamento è reciprocamente indipendente, che i loro livelli hanno effetti congiuntamente additivi, che sono sufficientemente densi o comunque equidistanti e che nessuno di loro è infinitamente piccolo rispetto agli altri.

Dati questi presupposti, allora per i valori a e b in A e m e n in M esistono delle funzioni a valori reali f su A e g su M, tali per cui [Krantz, Luce, Suppes e Tverski 1971]

$$(b, m) \geq (a, n) \Leftrightarrow f(b) + g(m) \geq f(a) + g(n) \qquad \text{(eq. 9.28)}$$

su cui si basa il teorema di rappresentazione della struttura empirica su quella formale dell'insieme dei numeri reali. Tali funzioni godono della proprietà di unicità (invarianza) rispetto a trasformazioni lineari, il che colloca la misurazione additiva congiunta al livello delle scale a intervalli. Date due altre funzioni f' e g', rappresentazioni alternative degli attributi A e M, allora esistono le costanti a valori reali α, β_a e β_m, con $\alpha > 0$, per le quali vale

$$f' = \alpha f' + \beta_a \qquad \text{(eq. 9.29)}$$

e

$$g' = \alpha g' + \beta_m \qquad \text{(eq. 9.30)}$$

9.7 Misurazione congiunta additiva: rivoluzione mancata?

Alle spalle dei risultati illustrati in modo sintetico in queste pagine, si erge una complessa struttura di teoremi e dimostrazioni algebriche, sviluppate nel tempo e raccolte in tre volumi sui *Foundations of Measurement* (Fondamenti della misurazione), opere collettive di Krantz, Luce, Suppes e Tversky, la cui ambizione è affrontare l'assiomatizzazione di tutte le possibili forme di misurazione sulla base della natura degli attributi. Il primo volume, *Additive and Polynomial Representations*, edito nel 1971,

riguarda tra le altre cose la misurazione additiva congiunta nella forma da noi analizzata. A distanza di quasi venti anni saranno pubblicati, vicini tra loro, il secondo e il terzo volume: *Geometrical. Threshold, and Probabilistic Representations* [Suppes, Krantz, Luce e Tversky 1989] e *Representation, Axiomatization, and Invariance* [Luce, Krantz, Suppes e Tversky 1990].

All'indomani della pubblicazione completa dei tre volumi, Cliff [1992] propose un bilancio dell'impatto della teoria assiomatica-rappresentazionale della misurazione, considerata un possibile grande passo avanti nello sviluppo dei modelli di misurazione in ambito psico-sociale. La rassegna di Cliff si conclude con un giudizio abbastanza netto sul fatto che l'impatto reale di questo promettente filone sarebbe stato molto sotto la sua potenziale portata[109], superato dal successo di approcci maggiormente orientati all'adattamento ai dati. La diffusa giustificazione di tipo pragmatico dei livelli di misurazione delle scale, riservata al solo giudizio dei ricercatori, e la scarsa attenzione data alla rimozione dell'interazione tra gli attributi, per aumentare la possibilità di raggiungere risultati migliori, sono due degli esempi che porta per mostrare la scarsa influenza della teoria assiomatica.

Cliff individua ed elenca alcuni nodi critici che avrebbero limitato la diffusione della teoria:

i. l'utilizzo di una *matematica complessa e astratta*, in buona parte estranea e troppo sofisticata per le conoscenze dei ricercatori e, in generale, di chi non abbia competenze specifiche nel campo dell'algebra;
ii. la *mancanza di una comprovata efficacia empirica*, ossia di esempi di ricerche di successo e grande richiamo, che sarebbero state capaci di attrarre l'attenzione di altri studiosi e di fungere da architrave paradigmatico per successivi passi lungo la strada della misurazione fondamentale assiomatica;
iii. il problema del *trattamento degli errori*, dato dalla natura deterministica degli assiomi, che non danno soluzioni su come interpretare e inglobare le deviazioni dal perfetto rispetto degli assunti;

[109] Cliff annuncia la sua conclusione già nel titolo dell'articolo: la teoria della misurazione astratta e la rivoluzione che non è mai avvenuta (*Abstract Measurement Theory and the Revolution That Never Happened*).

iv. gli *stili di ricerca* nell'ambito degli studi sugli atteggiamenti, generalmente rivolti all'analisi della varianza, con variabili multicategoriali anziché continue, alla ricerca delle interazioni tra attributi piuttosto che la loro eliminazione;

v. la preferenza per altre strade[110], apparentemente foriere di *sviluppi* promettenti, con una minore complessità e un minor dispendio di energie, che hanno "distratto" attenzione e risorse.

In un campo particolare come quello della ricerca sociale e psicologica, in cui lo sviluppo teorico della definizione di costrutti e concetti riguardanti l'oggetto di ricerca ha difficoltà molto maggiori rispetto ad altri ambiti, data la natura contemporaneamente non estensiva e non manipolabile di molti di essi, non è arduo immaginare il tipo di ostacoli che possa incontrare un approccio come quello assiomatico, per quanto sia potente e raffinato nel dimostrare i caratteri profondi della struttura di attributi. Nelle condizioni di ricerca date, soprattutto i punti i), iii) e iv) possono costituire problemi di non poca portata.

Individuare in un modello alternativo le caratteristiche della misurazione additiva congiunta, superando magari le difficoltà tecniche e le rigidità del sistema assiomatico, vorrebbe dire accogliere i vantaggi di diversi approcci. Il modello di Rasch ha le caratteristiche per candidarsi, in questo senso, a rappresentare un punto d'incontro. La discussione che segue è intesa a ricostruire una parte del dibattito su questo tema, provando a fornirne una sistemazione e alcuni contributi.

[110] Cliff porta come esempi il lavoro di Sternberg [1969] sul fronte sperimentale e quello di Jöreskog [1969] su quello correlazionale.

10 Modello di Rasch e ACM: gli elementi di affinità

Ci sono diversi, fondati motivi per cui il modello di Rasch può essere interpretato come un caso di misurazione congiunta additiva: più precisamente, come un'istanza *probabilistica* di ACM. Modello di Rasch (in generale la IRT) e ACM hanno camminato su strade parallele, sono cioè frutto di genesi distinte e di approcci tra loro differenti, nati e cresciuti in reciproca autonomia. Ben presto, diversi studiosi hanno cominciato a rilevarne le affinità[111]. L'interesse principale è presto detto: provare a dimostrare che un modello relativamente ben maneggiabile matematicamente e di facile comprensione, come quello di Rasch, può allo stesso tempo contenere e soddisfare gli assiomi della ACM. Questo punto, quello pratico, ci sembra il più importante di tutti. Del resto, prima ancora di Cliff, Falmagne [1976; cit. in Perline, Wright, Weiner 1979, 237] sottolineava che

> nel loro corrente *status*, le teorie sulla misurazione fondamentale sono algebriche, cioè, deterministiche. Le loro previsioni non si prestano facilmente alla verifica empirica. Qualsiasi allontanamento dei dati dalla teoria comporta un rebus verso il quale non si applicano le comuni regole di decisione della statistica.

Scrive Brogden [1977, 632-3] che "se uno è disponibile ad accettare p_{ia} [la probabilità di *a* di superare *i*, N.d.A.] come una misura ordinale dell'effetto congiunto della difficoltà dell'*item* e dell'abilità di una persona, può essere mostrato che il modello di Rasch è un caso speciale di misurazione congiunta additiva". In questo senso, il modello di Rasch si presenta come una "realizzazione pratica" [Perline, Wright e Weiner 1979, 237] di misurazione congiunta[112].

[111] A titolo riassuntivo, citiamo Keats [1967], Fischer [1968], Brogden [1977], Perline, Wright e Weiner [1979], Green [1986], Embretson e Reise [2000], Karabatsos [2001], Boorsboom e Mellenbergh [2004], Boorsboom e Zand Scholten [2008], Burro [2009], Zand Scholten [2011].

[112] "Poiché i dati educazionali e psicologici non sono generalmente perfettamente affidabili, l'assenza di una teoria dell'errore ha limitato l'utilità dei modelli di misurazione congiunta. Una forte eccezione è l'utilizzo diffuso dei modelli di Rasch" [Green 1986, 141]; "Solamente il modello di Rasch soddisfa pie-

Il grado nel quale i due approcci siano tra loro sovrapponibili è ovviamente sottoposto a sfumature di giudizio e, anche tra chi riconosce una certa affinità strutturale, alcuni tengono comunque a individuare punti di distanza formali [es. Karabatsos 2001] o ben più profonde distinzioni strutturali ed epistemologiche [es. Kyngdon 2008a; Michell 2008b].

Andiamo ora a rileggere le caratteristiche del modello di Rasch, attraverso le categorie della ACM.

Le ipotesi del modello, lo ricordiamo, prevedono l'esistenza di dimensioni latenti quantitative e continue, che possono essere costituite da attributi di vario tipo, per esempio atteggiamenti, dimensioni valoriali, abilità ecc.; si suppone, inoltre, che determinati soggetti e *item* possiedano un certo livello dell'attributo in questione e che, sulla base di quello, possano essere ordinati lungo il *continuum* latente. L'ordinamento avviene in modo simultaneo, a partire dalle risposte dei soggetti agli *item*, in modo da stabilire congiuntamente il livello di "abilità" dei primi e di "difficoltà" dei secondi. Per stimare le posizioni degli elementi su un'unica scala con un livello di misurazione intervallare, si assume che la probabilità dei soggetti (date le proprie abilità) di eguagliare gli *item* (date le loro difficoltà), sia definita da una funzione di probabilità parametrica basata sull'effetto additivo congiunto dell'abilità e della difficoltà. Queste ultime, sotto forma di parametri, sono stimate per ogni individuo e ogni affermazione in modo da massimizzare il più possibile l'adattamento con le effettive risposte registrate.

Ciò che si ottiene è rappresentato nella tabella 10.1: una matrice con i soggetti in riga, gli *item* in colonna e le rispettive probabilità congiunte nelle celle. Gli assunti di unidimensionalità e monotonicità del modello, garantiscono che i soggetti e gli *item* possano essere ordinati in modo crescente o decrescente.

namente le condizioni della misurazione congiunta e quindi è spesso preferito nelle applicazioni dove le proprietà della scala di misurazione sono ritenute molto importanti" [Embretson e Reise 2000, 149-50]; "La costruzione di una variabile latente utilizzando l'analisi dell'*item* di Rasch non è niente di meno che l'esame empirico della struttura quantitativa di quella variabile latente" [Barret 2003, 429].

		item							
		1	2	3	…	…	…	*n*	
soggetti	1	p_{11}	p_{12}	p_{13}	…	…	…	p_{1n}	$p_{1.}$
	2	p_{21}	p_{22}	p_{23}	…	…	…	p_{2n}	$p_{2.}$
	3	p_{31}	p_{32}	p_{33}	…	…	…	p_{3n}	$p_{3.}$
	…	…	…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…	…
	N	p_{N1}	p_{N2}	p_{N3}	…	…	…	p_{Nn}	$p_{N.}$
		$p_{.1}$	$p_{.2}$	$p_{.3}$	…	…	…	$p_{.n}$	

Tab. 10.1 – Matrice "soggetti x *item*" con le relative probabilità nel modello di Rasch

Le funzioni/curve di risposta all'*item* (vettori-colonna) e di risposta dei soggetti (vettori-riga) sono strettamente crescenti e tra loro parallele, cioè non s'intersecano mai. Ciò si rispecchia nell'invarianza dell'ordine delle probabilità in ogni riga e in ogni colonna: l'ordine globale dei soggetti non cambia anche se considerato condizionatamente a ogni singolo livello degli *item*, e viceversa per gli *item* condizionatamente ai livelli dei soggetti. In parole povere, l'ordine complessivo dei soggetti e degli *item* si mantiene in ogni vettore-riga e in ogni vettore-colonna. Questo significa che le colonne e le righe possono essere permutate in modo da ottenere un andamento monotonico in qualsiasi entrata della matrice.

10.1 Il confronto degli assunti

A questo punto, analizziamo gli assiomi della ACM, controllando punto per punto se gli assunti e le proprietà del modello di Rasch riescono a "coprirne" tutti i vari aspetti [Karabatsos 2001; Zand Scholten 2011]. Occorre ricordare nuovamente un fatto importante: in questo contesto non stiamo tenendo conto della questione dell'adattamento ai

dati empirici, ma solo delle caratteristiche formali del modello e, quindi, del caso-limite in cui i dati raccolti siano perfettamente coincidenti con quelli riprodotti dal modello.

Osserviamo, dunque, una matrice fittizia (tab. 10.2), contenente in riga una serie di livelli di abilità, in ordine decrescente; in colonna, una serie crescente di livelli di difficoltà, interpretati però come livelli decrescenti di "facilità" per rendere più immediata l'intuizione della struttura additiva. Possiamo, infatti, definire la facilità di un *item* come l'inverso della sua difficoltà. Pertanto, se il parametro di difficoltà è rappresentato da δ, quello di facilità, simboleggiato da τ, sarà pari a $-\delta = \tau$. Pertanto, essendo la funzione di probabilità governata dalla relazione $(\beta - \delta)$, sostituendo si ottiene $(\beta - \delta) = (\beta + \tau)$. I valori delle celle sono determinati dalla funzione di probabilità

$$\frac{1}{1+e^{(\beta-\delta_i)}} = \frac{1}{1+e^{(\beta+\tau_i)}} \qquad \text{(eq. 10.1)}$$

facilità dell'*item*

abilità del soggetto	**0,5**	**0,4**	**0,3**	**0,2**	**0,1**	**0**	**-0,1**	**-0,2**	**-0,3**	**-0,4**	**-0,5**
0,5	0,73	0,71	0,69	0,67	0,65	0,62	0,60	0,57	0,55	0,52	0,50
0,4	0,71	0,69	0,67	0,65	0,62	0,60	0,57	0,55	0,52	0,50	0,48
0,3	0,69	0,67	0,65	0,62	0,60	0,57	0,55	0,52	0,50	0,48	0,45
0,2	0,67	0,65	0,62	0,60	0,57	0,55	0,52	0,50	0,48	0,45	0,43
0,1	0,65	0,62	0,60	0,57	0,55	0,52	0,50	0,48	0,45	0,43	0,40
0	0,62	0,60	0,57	0,55	0,52	0,50	0,48	0,45	0,43	0,40	0,38
-0,1	0,60	0,57	0,55	0,52	0,50	0,48	0,45	0,43	0,40	0,38	0,35
-0,2	0,57	0,55	0,52	0,50	0,48	0,45	0,43	0,40	0,38	0,35	0,33
-0,3	0,55	0,52	0,50	0,48	0,45	0,43	0,40	0,38	0,35	0,33	0,31
-0,4	0,52	0,50	0,48	0,45	0,43	0,40	0,38	0,35	0,33	0,31	0,29
-0,5	0,50	0,48	0,45	0,43	0,40	0,38	0,35	0,33	0,31	0,29	0,27

Tab. 10.2 – Matrice "abilità x facilità" con le relative probabilità stimate secondo il modello di Rasch

Arriviamo, dunque, alla discussione degli assiomi vera e propria. Si ricordi che ogni punto illustrato nei seguenti paragrafi può essere controllato direttamente anche osservando la tab. 10.2.

Il primo assioma è quello dell'ordine parziale. Esso è soddisfatto dal modello di Rasch. Infatti, le probabilità stimate dal modello crescono monotonicamente rispetto all'abilità dei soggetti e alla facilità degli *item*[113]. A parità di *item*, la probabilità di superarlo cresce solamente se aumenta anche l'abilità dei soggetti. Ciò verifica la condizione dell'ordine parziale, secondo la quale dati due soggetti, s_1 e s_2, e un *item*, i_1, se $(s_1,i_1) \geq (s_2,i_1)$, allora $s_1 \geq s_2$. Inoltre, anche le condizioni di transitività e connessione sono soddisfatte, perché se $s_1 \geq s_2$ e $s_2 \geq s_3$, allora anche l'abilità del soggetto s_1 sarà maggiore o uguale a quella di s_3 e perché qualsiasi soggetto è confrontabile con qualsiasi affermazione posta in colonna.

Il secondo assioma è quello d'indipendenza, o di cancellazione singola. Anche questo è soddisfatto dal modello di Rasch. Riflettiamo sulle curve di probabilità: esse sono strettamente crescenti e tra loro parallele. Questo vuol dire che un *item* più difficile avrà sempre una probabilità minore di essere superato rispetto a uno più facile, qualsiasi sia il livello di abilità selezionato; graficamente ciò risulta nel fatto che la curva di risposta al primo *item* sarà costantemente al di sotto di quella del secondo. Lo stesso ragionamento vale per l'abilità: un soggetto con un grado maggiore di proprietà avrà sempre una probabilità maggiore di superare qualsivoglia *item* rispetto a un soggetto con meno proprietà, a prescindere dal livello dell'*item*. Si noterà che il contenuto di questo assioma trova una corrispondenza nel modello di Rasch nella proprietà dell'oggettività specifica, che afferma appunto che la posizione dei soggetti, come quella degli *item*, è indipendente, rispettivamente, dagli *item* o dai soggetti scelti per il confronto.

Passiamo ora all'assioma di cancellazione doppia. In questo caso è utile osservare prima la tabella 10.2 per avere un saggio del fatto che anche questa complessa serie di

[113] Il che, si noti, è equivalente a dire che la probabilità cresce monotonicamente rispetto alla differenza tra abilità e difficoltà.

relazioni d'ordine è necessariamente soddisfatta dal modello di Rasch[114]. Anche in questo caso le ragioni del successo risiedono nella natura delle funzioni di probabilità, che descrivono curve continue e costantemente crescenti, tra di loro ordinate e parallele. Questo carattere si rintraccia in matrice, dove i vettori-riga e colonna contengono i valori delle probabilità, che si dispongono in modo da rendere molto bene l'idea di curve parallele e traslate, perché si presentano come il frutto di uno "scorrimento" in diagonale di un medesimo vettore ripetuto: aumentando la densità dei livelli, percorrendo le diagonali in direzione 1 e 2 troveremmo comunque rispettati i criteri che costituiscono il principio della doppia cancellazione.

Come si è già segnalato, gli assiomi di risolvibilità e condizione archimedea non sono empiricamente testabili su insiemi finiti di oggetti, elementi, livelli ecc. Un modo di sostituire un test diretto con una prova indiretta è la verifica della gerarchia dei gradi di cancellazione ammessi dalle dimensioni della matrice [Scott 1964]. In una condizione ideale, come quella rappresentata dalle stime delle probabilità del modello di Rasch, anche la gerarchia completa si dimostra accettata, come la tabella 10.2 mostra in via teorica.

Una prova indiretta dell'assioma di risolvibilità è possibile, anche se solo in via di principio, se si è disposti ad accettare quel procedimento che Galileo definì "esperimento mentale". Infatti, data una certa probabilità x che un soggetto s_1 superi un item i_1, allora dato un soggetto s_2 con un'abilità inferiore a s_1, è possibile *immaginare* che possa esistere sempre un *item* i_2 meno difficile di i_1 e tale per cui la probabilità di s_2 di superarlo sia uguale a x. Lo stesso ragionamento vale, *mutatis mutandis*, per la ricerca di un soggetto s_2 dati s_1, i_1 e i_2. In sostanza, l'esperimento mentale si riduce alla possibilità che tra due qualsiasi soggetti e due qualsiasi *item* si possa sempre trovare un terzo soggetto e un terzo *item* le cui abilità e difficoltà si collochino in un punto intermedio tra le prime due. L'esperimento, nell'ambito di certi limiti, ha una sua applicabilità empirica, ma se portato alle sue conseguenze ultime è solo virtuale perché da un punto di vista pratico e strumentale è ovviamente impossibile procedere all'infinito nella ricerca di soggetti e *item* con posizioni sempre più ravvicinate tra loro.

[114] Questo, in realtà, vale per l'intera gerarchia di cancellazioni, come si può facilmente controllare nella tabella 10.2. Si veda anche oltre nel testo.

Per quanto riguarda la condizione archimedea, il modello di Rasch rispetta il requisito formale per cui se un primo soggetto, con una certa abilità, ha una probabilità data di superare un *item*, allora un soggetto con un'abilità maggiore avrà la stessa probabilità di superare un *item* più difficile (cioè, con una "facilità" minore).

Quanto fin qui esposto, giustificherebbe la legittimità dell'ipotesi che il modello di Rasch sia un caso di misurazione congiunta additiva di abilità e difficoltà attraverso le relazioni d'ordine tra le probabilità.

Un altro modo di interpretare tale affinità è quello dell'analogia con la misurazione derivata in fisica [Rasch 1960; Andrich 1988; Fischer 1995b; per una visione critica, Kyngdon 2008a]. Luce [1987] mostra che alcune istanze di misurazione derivata sono esempi di misurazione congiunta, caratterizzate da regole di combinazione di tipo moltiplicativo. E' il caso della seconda legge del moto di Newton, per la quale l'accelerazione è il rapporto tra forza e massa di un corpo ($A = F/M$). A parità di forza, il rapporto delle accelerazioni di due corpi si risolve nel rapporto inverso tra le masse: tale rapporto è quindi indipendente dalla forza, che sparisce dall'equazione. Una trasformazione logaritmica muta il rapporto tra le masse in differenza tra i loro logaritmi.

$$\frac{A_1}{A_2} = \frac{\frac{F}{M_1}}{\frac{F}{M_2}} = \frac{M_2}{M_1} = \ln(M_2) - \ln(M_1) \qquad \text{(eq. 10.2)}$$

Riprendiamo ora la derivazione del modello di Rasch, illustrata nel secondo capitolo. Come si ricorderà, in quel caso il rapporto tra abilità e difficoltà è trasformato in differenza tra i loro logaritmi. La proprietà di oggettività specifica comporta che, partendo dai *logit* delle probabilità, il rapporto tra le abilità dei soggetti (o la differenza tra i loro logaritmi) siano indipendenti dalle difficoltà degli *item*. Viceversa per quanto riguarda le difficoltà degli *item* rispetto alle abilità dei soggetti.

$$\frac{\left[\frac{P(X_{pj}=1}{P(X_{pj}=0}\right]}{\left[\frac{P(X_{qj}=1}{P(X_{qj}=0}\right]} = \beta_p - \beta_q = \ln\left(A_p\right) - \ln\left(A_q\right) = A_p/A_q \quad \text{(eq. 10.3)}$$

Su questa base, Fischer [1995b, 31] conclude che "sebbene le scale originali B e D (abilità e difficoltà, N.d.A.) sono solamente ordinali per la natura della nozione psicologica di "abilità" e "difficoltà", le trasformazioni sono altamente specifiche e uniche rispetto a trasformazioni lineari" e questo renderebbe il caso della misurazione derivata in fisica e quello, sotto nostro esame, del modello di Rasch, tra loro analoghi. Attraverso questo parallelo, si potrebbe giustificare in altro modo il fatto che il modello di Rasch sia un caso di misurazione moltiplicativa congiunta, con una struttura additiva dopo una trasformazione logaritmica.

Nel prosieguo del testo, analizzando le critiche all'ipotesi che il modello di Rasch sia un caso di ACM, lasceremo da parte quest'ultima interpretazione, seppur formalmente molto interessante: poiché la discussione si concentrerà sulla condizione preliminare della natura quantitativa degli attributi e sul fatto che il modello di Rasch possa aiutare o no a indagarla, l'ipotesi di una misurazione derivata risulterà necessariamente secondaria se prima non troverà risposta la questione dello *status* degli attributi da cui discenderebbe.

11 Modello di Rasch e ACM: gli elementi di differenza

Nel precedente capitolo abbiamo cercato di raccogliere gli elementi formali che suffragano l'ipotesi di una stretta parentela tra il modello di Rasch e la ACM; ora è però necessario analizzare, al contrario, le ragioni strutturali ed epistemologiche che spingono un altro schieramento di studiosi di scienze psicologiche e sociali a rifiutare tale accostamento. Prima di entrare nel dibattito e provare a darne una ricostruzione tematica, è utile affrontare alcune differenze relativamente più intuitive che distinguono le due famiglie di modelli.

11.1 Determinismo *versus* probabilità

Il primo aspetto, fondamentale, riguarda la natura dei modelli IRT e di quelli ACM. Mentre i primi sono probabilistici, gli assiomi della misurazione congiunta additiva hanno un carattere strettamente deterministico, di stampo algebrico[115]. Sul versante della ACM, questo significa che un rigoroso controllo degli assiomi comporta un rifiuto in blocco del modello anche in presenza, al limite, di un solo caso divergente dalle ipotesi. Le violazioni delle condizioni degli assiomi sono considerate "errori" che negano la possibilità che la struttura indagata sia quantitativa continua. Certo, si potrebbe discutere sull'appropriatezza di un termine come "errore", che pare più adatto se utilizzato con riferimento a concetti come quello di "validità" (per cui, per esempio, non si rileva in modo corretto una certa proprietà) più che per descrivere la non riproducibilità di un dato empirico da parte di un modello o un assunto formale astratto. Se una certa struttura empirica non possiede le caratteristiche richieste da una teoria della misurazione, allora i suoi elementi non sono "errati": semplicemente violano i presupposti necessari a darne una certa rappresentazione.

Riguardo alla natura deterministica della ACM si può ripetere la constatazione svolta per altri modelli deterministici nella Parte Prima: nella pratica della ricerca psico-

[115] Questo, come detto, comporta che alcuni assiomi si muovano nell'ambito degli insiemi infiniti di elementi. Per questo motivo si utilizzano strade alternative, quale la gerarchia dei gradi di cancellazione.

sociale è raro, se non quasi impossibile, che una serie di dati empirici rispetti in modo integrale e completo un insieme di assunti che prevedono, in modo rigido e inflessibile, il rispetto di certe condizioni formali. Questo accade con modelli di tipo ordinale, quale quello di Guttman, e accade prevedibilmente a maggior ragione con modelli che amplificano la selettività dei requisiti. Ora, se una deviazione dagli assiomi è realmente frutto di uno o più errori *casuali*, allora questo non implica necessariamente e per principio che gli attributi in questione non rispettino gli assiomi. Il problema è che un modello deterministico non è in grado di gestire, in nessun caso, queste deviazioni e non offre strumenti per valutarne l'eventuale livello di accettabilità. A questo si riferiscono gli autori citati in precedenza, quando lamentano una scarsa applicabilità degli assunti ACM e la totale mancanza di flessibilità. Del resto, per sfruttare pienamente la potenza della struttura algebrica degli assiomi, questa rigidità è il prezzo da pagare.

Poniamo, però, che le violazioni degli assiomi non scaturiscano da errori casuali, ma rappresentino vere e proprie deviazioni della struttura empirica da quanto immaginato da quella formale del modello. Nel nostro caso specifico, ciò può significare che gli attributi non hanno effettivamente niente a che vedere con strutture quantitative, combinabili additivamente in modo congiunto: cioè, non esiste nessuna dimensione latente che influenza le risposte delle persone a una serie di affermazioni, oppure non ne esiste una rappresentazione additiva perché abilità e difficoltà non hanno una struttura abbastanza complessa da sostenerla.

C'è però un'altra possibilità, contemplata dai modelli come quello di Rasch, che sostiene quanto segue: l'ipotesi del costrutto latente è sostenibile ed è possibile rappresentare su di esso abilità e difficoltà, utilizzando una loro combinazione additiva; solamente, la relazione tra loro si basa su un rapporto probabilistico. L'atteggiamento di una persona verso un oggetto cognitivo può avvicinarsi più o meno a quello espresso da una certa affermazione, ma alla fine il soggetto deve "tirare una riga", accettare o rifiutare un *item*, valutare una risposta corretta ecc., *dato che il modello tratta solo chiusure dicotomiche*[116]. Per questo si deve tenere conto di un "effetto alone" nelle risposte, che non è necessariamente frutto di errori di valutazione, ma l'emergere del fatto che gli atteggiamenti sono composti da una complessa pluralità di gradi, che la distanza tra posi-

[116] Questo ovviamente vale, in misura maggiore o minore, per tutte le tecniche a risposta chiusa. Tale processo è meno problematico per quanto riguarda il dare una risposta giusta o sbagliata su precise nozioni.

zione personale ed espressione dell'*item* si allarga o si restringe e che, quando queste soglie si avvicinano molto, a far pendere la risposta da una parte o dall'altra possono essere sottili sfumature, il prevalere contingente di questo o quell'aspetto del costrutto. Il vantaggio dei modelli probabilistici è quello di tenere conto di questa variabilità, di questo "alone" e del fatto che, pur tracciando una linea netta tra un "sì" e un "no", un soggetto non sempre accetta o rifiuta *in toto* un'affermazione, possono esserci gradi diversi che si rispecchiano in vari livelli di probabilità[117]. In pratica, il modello di Rasch (e gli altri modelli IRT) di fronte alla condizione pratica delle scienze umane, offre un'interpretazione di questo "effetto alone" e gli strumenti per valutarne un possibile impatto, a scapito dell'assoluto, tetragono rigore assiomatico dei modelli ACM. Su questo punto si svolge un pezzo del dibattito sull'adeguatezza del modello di Rasch come istanza di ACM, tra chi rivendica un maggior realismo del primo [es. Borsboom e Mellenbergh 2004] e chi invece ritiene che l'elemento stocastico stesso dimostri l'impossibilità di provare la natura quantitativa dei costrutti mentali [es. Kyngdon 2008b].

11.2 Assunti ACM e modello di Rasch: un legame tra strutture formali

C'è un altro aspetto importante, con molteplici risvolti, che non ci siamo mai stancati di ripetere nel corso del testo. Le proprietà che rendono il modello di Rasch uno strumento così potente e duttile, appartengono al modello formale stesso. Lo abbiamo notato per l'invarianza, per l'oggettività specifica, per le statistiche sufficienti e lo abbiamo fatto anche per la rispondenza agli assiomi della ACM: tutti elementi e caratteristiche desiderabili e ambite per garantire progressi nella ricerca delle scienze umane, ma che riguardano integralmente la struttura matematica del modello, oppure al limite dati rilevati empiricamente che si adattino perfettamente a esso. Questo significa che non

[117] Scrive Marradi [2007, 155], a proposito del modello di Rasch, che esso "trasforma in complessi procedimenti matematici l'intuizione di base che non ci si può attendere che un soggetto approvi inevitabilmente tutte le frasi più "facili" della sua posizione in una dimensione e disapprovi inevitabilmente tutte le frasi più "difficili": si può considerare solo la probabilità che questo si produca". Lo studioso conclude, ironicamente, che dal suo punto di vista "tutto ciò è giusto, ma applicare i calcoli sofisticati di Rasch a risposte che quasi sempre i soggetti danno in modo approssimativo e superficiale è come sparare a una farfalla con un cannone".

basta stimare i parametri del modello su dati empirici per estendere a questi ultimi le sue pregevoli caratteristiche. Esso produce delle stime, quelle più verosimili, più vicine possibili alle distribuzioni di frequenza delle risposte. E produce delle stime da *qualsiasi* base empirica di risposte dicotomiche di un campione di soggetti a un *pool* di *item*. È compito successivo del ricercatore dare una risposta alla domanda "quanto vicine sono le stime ai dati di partenza?". Il modello offre i suoi *output* in ogni caso, anche quando l'analisi successiva dimostra che l'adattamento è del tutto inadeguato e le stime completamente inaffidabili, anche quando cioè il buonsenso del ricercatore lo porta a concludere di trovarsi di fronte a una base empirica che nega alla radice gli assunti del modello.

Tutto ciò ci ricorda che mentre le stime, confacendosi alle funzioni logistiche di probabilità, rispettano sempre in via teorica gli assiomi della ACM, i dati reali alla loro origine possono invece essere in parte o del tutto incoerenti con essi.

Karabatsos [2001] mette inoltre in luce il fatto che nemmeno l'attestazione dell'adattamento può ritenersi immune dal rischio di distorsioni. Difatti, si potrebbe essere indotti a pensare che, poiché i valori stimati dal modello, in quanto prodotti dalla funzione logistica, aderiscono necessariamente agli assiomi della ACM, allora il *fit* dei dati al modello debba essere sempre una misura del grado di adesione del materiale empirico agli assunti. Ma come mostrato già da Nickerson e McClelland [1984; cit. in Karabatsos 2001] si danno casi in cui un modello di misurazione congiunta numerico[118] si adatta perfettamente a dati che tuttavia violano gli assiomi in modo consistente. Questo apparente paradosso è reso possibile dal fatto che, come ogni modello dipendente dai dati, anche in quello di Rasch la stima dei parametri tende ad "assorbire" ogni forma (casuale o sistematica) di disturbo (*noise, disturbance*) nei dati. Di nuovo, come il termine "errore" anche quello "disturbo" può dare, dal nostro punto di vista, una visione fuorviante: una struttura empirica che non segue gli assiomi della ACM o, in generale, non presenta caratteri quantitativi o ordinali, non è di per sé "disturbata" o "errata". Quello che però vale in generale è che nel processo di stima dei parametri, qualsiasi carattere dei dati deviante dalla struttura formale avanzata dal modello sarà tendenzial-

[118] Karabatsos [2001] fa una distinzione, sostenendo che "entrambi, i modelli di Rasch e gli assiomi di misurazione congiunta, specificano IRF parallele, tuttavia ognuno utilizza un approccio differente. Mentre il modello di Rasch specifica IRF parallele usando una funzione numerica per limitare P_{ng}, la teoria della misurazione congiunta definisce la forma delle IRF parallele con limitazioni d'ordine non numeriche su P_{ng}."

mente "mimato", riprodotto, assorbito appunto nella stima, che per definizione minimizza i residui tra osservazioni e probabilità stimate. Per questo, la quantificazione dell'adattamento avviene tra i dati reali e una stima che è dipendente da quei dati, producendo un "effetto di mascheramento" che sistematicamente sopravvaluta il livello di *fit*[119].

Sul versante della ACM, invece, gli assiomi sono riparati da simili rischi, perché avanzano requisiti non condizionati in nessun modo da alcun *set* di dati, perciò sono completamente indipendenti da fattori interni alla struttura empirica stessa.

11.3 ACM, strutture empiriche e attributi mentali: un nodo critico

C'è infine (mai come in questo caso "*last but not least*") un aspetto fondamentale che appartiene alla natura delle proprietà psicologiche: non sono proprietà estensive, direttamente tangibili su qualche oggetto o manipolabili, a differenza della lunghezza o del peso, e quindi è più difficile teorizzarne la struttura. Quando Zand Scholten [2011, 40-41] tratta tre aspetti di criticità nel confronto tra modello di Rasch e ACM, sta di fatto concentrando la sua attenzione su tre effetti di questo stesso aspetto epistemologico.

Il primo punto riguarda il fatto che, nell'ambito della Teoria Rappresentazionale della Misurazione da cui è nata la Teoria ACM, si ritiene necessario specificare *a priori*, in modo preciso e non equivoco, a quale insieme di oggetti gli assiomi dovrebbero applicarsi. Se questo compito è relativamente accessibile rispetto alla definizione di molte proprietà fisiche, "perfino per proprietà psicologiche apparentemente "semplici" quale la difficoltà di *item* che accertano l'abilità aritmetica o spaziale, questo è già molto più difficile" [*ibidem*].

Qui s'innesta il secondo punto, che ci porta di fronte al tema del grado di robustezza delle teorie e delle ipotesi concernenti le dimensioni attitudinali. Michell [2008a,

[119] Sempre Karabatsos [2001, 395] argomenta che "si può essere tentati di concludere da ciò che forse dovrebbero essere impiegate altre statistiche di adattamento, non basate sui residui, per testare l'accordo dei dati con il modello di Rasch. Tuttavia, le statistiche di adattamento non basate sui residui possono soffrire allo stesso modo il mascheramento. Qualsiasi statistica di adattamento basata sui parametri stimati β e δ assume che essi siano i veri valori dei parametri, non viziati da dati potenzialmente disturbanti."

15] ricorda quanto affermava molto tempo fa Boring [1920, 33], secondo il quale "non ha senso provare a ricercare nel processo logico di elaborazione matematica una precisione significativa per la psicologia che non sia [già] presente nella configurazione psicologica del problema". Detto rovesciando i termini, l'insieme dei concetti e delle loro relazioni, che compongono le ipotesi iniziali riguardo a un certo costrutto, dovrebbero essere abbastanza robusti da "tenersi insieme", anche in sede di analisi del modello, se l'obiettivo è quello della misurazione. Ora, di norma un ricercatore sociale parte, sulla base della propria conoscenza, definendo un concetto in grado di connotare un costrutto latente, descrivendone gli aspetti e caratterizzandolo in ogni suo elemento. Sulla base di questo lavoro, poiché il costrutto non è direttamente rilevabile, costruisce e seleziona un *set* di affermazioni che, per fondate e motivate ragioni teoriche, ritiene descrivano i vari livelli di espressione del costrutto, secondo una logica nel nostro caso cumulativa; dall'altro lato, seleziona un campione di soggetti che pensa abbiano una posizione lungo la dimensione definita dal costrutto. A questo punto, le risposte dei soggetti agli *item* sono trasformate dalla funzione del modello, così da arrivare alle stime. Tecnicamente è possibile, e in concreto accade spesso, che *item* o soggetti siano eliminati dalla considerazione del modello perché il loro livello di adattamento non è ritenuto adeguato, quindi per migliorare il *fit* globale del modello stesso. Così facendo si stanno modificando i rapporti semantici d'indicazione [Marradi 1980; 2007] tra il costrutto e gli indicatori inizialmente individuati: si concede, di fatto, che il legame ipotizzato tra concetto indicato e alcuni indicatori non sia valido, sulla base dell'osservazione empirica (e della sua interpretazione da parte del modello). Possono non esserci, quindi, ragioni di contenuto che spingano a espellere l'indicatore dal modello. Questo però, consapevolmente o meno, comporta che verosimilmente cambi anche il concetto teorico di partenza, perché evidentemente vi erano all'inizio delle componenti estranee di cui il processo di analisi ha mostrato la non omogeneità. Insomma, se alcuni indicatori spariscono, cambia in qualche modo la dimensione latente scalata dal modello. Esistono però valide motivazioni teoriche per giustificare la perdita di certi indicatori? E come cambia il costrutto iniziale? Zand Scholten [2011, 40], pur da un punto di vista non rappresentazionale, molto correttamente avverte che "quando il *focus* è solamente nella rivendicazione di un livello di misurazione intervallare per una qualche proprietà, e non nell'identificazione di fattori che possono confondere la variabile d'interesse, l'utilizzo del modello di Ra-

sch può darci un falso senso di successo e può condurre a strumenti di misurazione che rivendicano in modo fallace di rappresentare una qualche proprietà quantitativa". Insomma, dal punto di vista teorico della RMT, un modello *in sede di misurazione* non deve servire a selezionare le variabili-indicatore che legano di più tra loro, ma a testarne la scalabilità e la natura quantitativa, per cui la loro selezione originaria deve risiedere su fondate ragioni teoriche: gli *item* non possono entrare e uscire dal modello come i passeggeri su un autobus. Nelle scienze fisiche, talvolta, è più agevole aggiustare il tiro sul versante della teoria e trovare le giustificazioni per l'inadeguatezza di uno strumento di misurazione. Zand Scholten [*ibidem*] spiega egregiamente il meccanismo con questo esempio:

> Supponiamo che stessimo per costruire uno strumento di misurazione per la lunghezza usando la comparazione a coppie di persone con aste di legno. Se avessimo rimosso le aste non conformi ai nostri assiomi di additività, saremmo seriamente a rischio di produrre uno strumento di misurazione che produce una misurazione a livello d'intervalli solamente per un sottogruppo degli oggetti per cui gli assiomi dovrebbero valere. Ora, se potessimo squalificare queste aste anomale dimostrando che erano tutte aste fatte da legno appena tagliato che era ancora molto flessibile, potremmo legittimamente adeguare la nostra demarcazione di questo insieme di oggetti.

Il terzo punto è già stato affrontato nella nostra discussione e riguarda il fatto che alcuni assiomi debbano poter essere testati su insiemi infiniti e continui. Si è detto che, oltre alle gerarchie di cancellazione, è possibile immaginare di poter scendere asintoticamente nell'infinitamente piccolo solo attraverso un esperimento mentale: tra due *item* o due soggetti, potrò sempre trovarne un terzo che si colloca tra loro, e così avanti. Se ciò in principio è possibile, nel caso di proprietà psicologiche non direttamente operativizzabili è un'impresa quasi disperata riuscire a trovare sempre il successivo terzo elemento. Pensando a un test sulle capacità aritmetiche di un campione di soggetti, quale *item*, si chiede Zand Scholten, giace a un livello di difficoltà tra quello di '1+1' e quello di '1+2'?

I punti discussi aiutano a costruire un quadro di riferimento generale sulla questione del modello di Rasch come forma di misurazione additiva congiunta. Abbiamo visto i motivi per cui il modello si qualifica matematicamente come una struttura di quel tipo, ma anche alcuni argomenti che tendono a indebolire questo legame, rendendo la questione più complessa e articolata. Proveremo adesso ad andare più in profondità, affrontando alcuni aspetti del dibattito epistemologico a nostro avviso salienti. Per farlo, discuteremo a partire dalle posizioni di alcuni studiosi che negano, con argomenti diversi e partendo da visioni differenti, che il modello di Rasch sia un caso di ACM.

12 Il modello di Rasch come misurazione congiunta: una critica dal versante rappresentazionale

La prima prospettiva dalla quale vogliamo cominciare questa disamina è quella di Kyngdon [2008a; 2008b], che cerca di dimostrare come, partendo proprio da un rigoroso punto di vista rappresentazionalista e quindi in piena adesione ai principi della ACM, il modello di Rasch non possa essere legittimamente rivendicato come un'istanza di misurazione additiva congiunta. Lo fa partendo con l'affermare i presupposti della teoria rappresentazionale contemporanea (RMT) e le condizioni che governano l'ipotesi della misurazione fondamentale di una struttura empirica, adottando un taglio per certi versi addirittura ontologico.

12.1 Presupposti della ACM: perché il modello di Rasch fallirebbe nel soddisfarli

Innanzitutto, l'insieme di oggetti, fatti, eventi, relazioni, caratteristiche, comportamenti direttamente osservabili del mondo naturale possono costituire *strutture a livello di superficie*. Oggetti ed eventi da una parte e le relazioni che intercorrono tra loro dall'altra, costituiscono due insiemi distinti che, nel loro complesso, formano una *struttura empirica relazionale* (fig. 12.1). Kyngdon sottolinea [2008a, 91] che "qualsiasi struttura empirica non è un insieme puro come un numero reale dal momento che gli elementi di una struttura empirica sono insiemi i cui membri sono oggetti, eventi o relazioni naturali"[120]. Quindi, seguendo quello che definisce come "Platonismo", separa nettamente gli insiemi di oggetti, eventi, attributi naturali, appartenenti al mondo sensibile e quindi collocati nello spazio e nel tempo, dagli insiemi puri dei numeri, astratti, che regnerebbero in una dimensione del tutto esterna al mondo naturale, fuori dallo spazio e dal tempo, elementi astratti.

[120] La definizione di Kyngdon richiama quella di Suppes e Zinnes [1963, 7] per i quali "un sistema relazionale empirico è un sistema relazionale il cui dominio è un insieme di entità identificabili, come pesi, persone, affermazioni attitudinali, o suoni".

L'*assiomatizzazione* consiste nella costruzione di una serie coerente di proposizioni che prescrivono come una struttura empirica relazionale si debba comportare per essere rappresentata su una *struttura relazionale numerica*. Come la struttura empirica è composta di una serie di elementi e dalle relazioni che li connettono, così la struttura numerica è composta di una serie di numeri e dalle relazioni tra essi intercorrenti. Solitamente, la struttura numerica è composta dall'insieme dei numeri reali positivi e dalle relazioni che li legano (fig. 12.1).

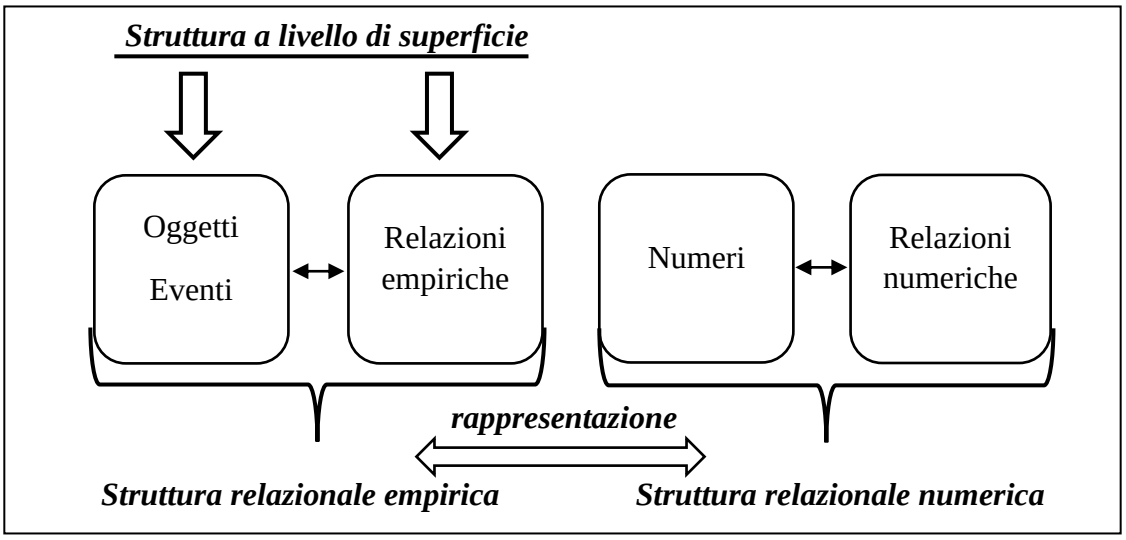


Figura 12.1 – Schema di una struttura relazionale empirica e del suo rapporto con una numerica

La struttura relazionale empirica, prosegue Kyngdon, è capace di sostenere una *rappresentazione* su quella numerica solamente se, attraverso il controllo degli assiomi sulla base empirica, si verifica l'esistenza di un *omomorfismo* tra le strutture, ossia se esiste un rispecchiamento tra i loro elementi e le relazioni. Ipotizzando che l'insieme degli elementi numerici sia quello dei numeri reali positivi, formalmente si scrive che[121]

$$\mathfrak{X} = \langle X, R \rangle \rightarrow \mathfrak{R} = \langle \mathbb{R}^{+}, S \rangle \qquad \text{(eq. 12.1)}$$

dove X è l'insieme degli elementi empirici

[121] Ogni tipo di omomorfismo riguarderà anche una o più *operazioni* ammesse sugli elementi degli insiemi, che ne preservino le relazioni. In questo caso si omettono dalle formule *solo* per semplificare e mantenere la generalità del ragionamento.

R è l'insieme di tutte le possibili relazioni di un certo tipo tra gli elementi di X

$\langle X, R\rangle$ è il dominio di $\mathfrak{X}$, la struttura relazionale empirica

$\mathbb{R}^+$è l'insieme di tutti i numeri reali positivi

S è l'insieme di tutte le possibili relazioni di un certo tipo tra gli elementi di $\mathbb{R}^+$

$\langle \mathbb{R}^+, S\rangle$ è il dominio di $\mathfrak{R}$, la struttura relazionale numerica.

Se esiste un omomorfismo che mappa gli elementi, le relazioni (e le operazioni ammesse) dell'insieme empirico su quello numerico, allora esisterà una funzione che descrive tale rapporto, espressa nel *teorema di rappresentazione*, e un'altra funzione, espressa nel *teorema di unicità*, che indica la famiglia di trasformazioni sulla funzione di rappresentazione che ne mantengono inalterata la mappatura[122].

All'interno di questa cornice teorica, il tracciato percorso da Luce e Tukey ha permesso di definire il concetto di additività nel modo più generale possibile, concedendogli di varcare le colonne d'Ercole che lo confinavano nell'ambito delle proprietà concatenabili.

Seguendo il ragionamento precedente sulle strutture empiriche, ipotizziamo che *A* e *M* siano due insiemi di oggetti, eventi, proprietà del mondo naturale, non direttamente manipolabili nel senso della concatenazione. Ammettiamo, poi, che esista un terzo insieme empirico *Z*, in relazione con i primi due, formato dal prodotto cartesiano degli elementi di *A* e *M*. La struttura relazionale empirica $\mathfrak{Z} = \langle A \times M, \geq\rangle$ si dice di *misurazione congiunta* se gli elementi di *Z* rispettano gli ormai ben noti assiomi rispetto alla relazione '≥'. Come conseguenza della validità della relazione '≥' sugli elementi di *Z*, allora esiste un teorema di rappresentazione e uno di unicità su entrambi *A* e *M*.

Interroghiamoci adesso sulla natura degli insiemi coinvolti nel modello di Rasch. Il primo, *B*, è l'insieme delle stime delle abilità di un campione di soggetti (β_1, β_2, β_3…), mentre il secondo, Δ, è l'insieme delle stime delle difficoltà di un gruppo di *item* (δ_1, δ_2, δ_3…). La terza variabile, *P*, è la probabilità dei soggetti di superare gli *item* ed è definita come una combinazione di *B* e Δ, costituita dalla funzione logistica in $P = f(B, \Delta)$. La

[122] Il teorema di unicità stabilisce anche, coerentemente con l'impostazione di Stevens, il livello di misurazione.

relazione '$\geq$' vale sempre sugli elementi di *P*, per la natura stessa della relazione (come si è mostrato in precedenza).

Ora, è possibile concludere che la struttura relazionale $\mathfrak{P} = \langle B \times \Delta, \geq \rangle$ sia dello stesso tipo di quella empirica di misurazione congiunta $\mathfrak{Z} = \langle A \times M, \geq \rangle$? Kyngdon [2008a, 99-100] argomenta che, secondo un piano di ragionamento rigorosamente rappresentazionale, la risposta sarebbe "no". Infatti, il tratto latente θ, sostiene, è definito come una porzione dell'insieme dei numeri reali, e come tale appartiene al mondo che egli chiama "degli insiemi puri"; poiché le stime numeriche contenute in *B* e Δ sono sottoinsiemi dei valori del tratto latente θ, allora anch'esse sono sottoinsiemi dell'insieme puro dei numeri reali. Si dovrebbe quindi concludere che né *B* né Δ sono composti di elementi collocati spazio-temporalmente, bensì da "insiemi puri di numeri reali del reame platonico [e quindi] sono sempre entrambi identici strutturalmente l'uno all'altro e agli altri insiemi puri di numeri reali, quali gli insiemi di probabilità" [*ibidem*].

L'unica conseguenza da trarre sarebbe quindi che $\mathfrak{P} = \langle B \times \Delta, \geq \rangle$ non è una struttura relazionale empirica di misurazione congiunta, perché il modello di Rasch mappa le probabilità, che sono numeri reali, su altri numeri reali, ossia gli insiemi delle stime dei soggetti e degli *item*. Una mappatura di una struttura relazionale numerica su un'altra struttura relazionale numerica non potrebbe intendersi in nessun caso come un'istanza di misurazione congiunta così come definita dalla RMT, perché mancherebbe un ingrediente fondamentale: una struttura empirica.

12.2 Commento all'argomento di Kyngdon

Da un punto di vista formale, se si accettano rigidamente le definizioni di Kyngdon dei sistemi relazionali, il suo argomento è ragionevolmente coerente. Si può comunque non condividere e mettere tra parentesi la sua posizione secondo cui le dimensioni latenti θ sarebbero parte, di per sé, del mondo degli insiemi puri, ossia prive di contenuto empirico: infatti, seppur non direttamente tangibili e manipolabili, esse possono essere concepite come la composizione della serie di tutte le possibili posizioni che un individuo può assumere nei confronti di un determinato oggetto cognitivo, comuni ai soggetti che condividono un insieme di riferimenti culturali e simbolici, manifestazioni

di gradi diversi nell'espressione di un certo atteggiamento. La questione non è se una dimensione latente, così intesa, *sia* un sottoinsieme dei numeri reali (come non lo è per alcun attributo che si manifesti tramite indicatori di qualsiasi tipo) ma se essa possa essere *rappresentata* sull'insieme dei numeri reali: proprio ciò che gli assunti della misurazione congiunta dovrebbero aiutare a chiarire. Se affermare l'esistenza di un qualsiasi ente fosse del tutto dipendente dal poterlo osservare e ciò che non si riuscisse a osservare, agli effetti dell'impresa scientifica, fosse da considerarsi non esistente nel dominio della realtà empirica, allora metteremmo seriamente nei guai anche molti altri ambiti scientifici, per esempio nella fisica (chi ha mai osservato direttamente una particella subatomica o un campo di forze?) Per cui, non ci pare la natura osservativa diretta di un fenomeno a marcare, in questo caso, il punto.

In ogni caso, sarebbe complicato negare che un essere umano è un elemento empirico, un "oggetto" di conoscenza appartenente al mondo della natura, come lo sono un'affermazione, una frase, un *item*: tutte cose che possiamo identificare e collocare nello spazio e nel tempo [Borsboom e Zand Scholten 2008]. Ed è un *fatto* che un soggetto, in un certo tempo e in un certo luogo, esprima la propria convinzione di ritenere accettabile o meno una certa affermazione (ferme restando le numerose questioni di validità ecc.) Ci pare quindi ragionevole sostenere che, fino a questo punto del processo cognitivo (ossia un soggetto che risponde a un *item*) ci troviamo in quello che Kyngdon definirebbe il "mondo della natura".

Quando poi dalle risposte verbali a una domanda si passa, attraverso un qualche modello matematico, a una formalizzazione delle relazioni osservate tra soggetti e *item*, mediante una serie di stime numeriche delle loro posizioni sulla dimensione che avrebbe ispirato le risposte, allora la dimensione latente θ è (eventualmente) rappresentata su un sottoinsieme dei numeri reali: in questo, un ulteriore sottoinsieme è composto dagli insiemi puri B e Δ, cioè le stime delle posizioni di soggetti e *item*. A questo punto non ci troveremmo comunque più, interpretando la logica di Kyngdon, nella sfera empirica ma in quella della rappresentazione numerica (il "mondo platonico") (fig. 12.2).

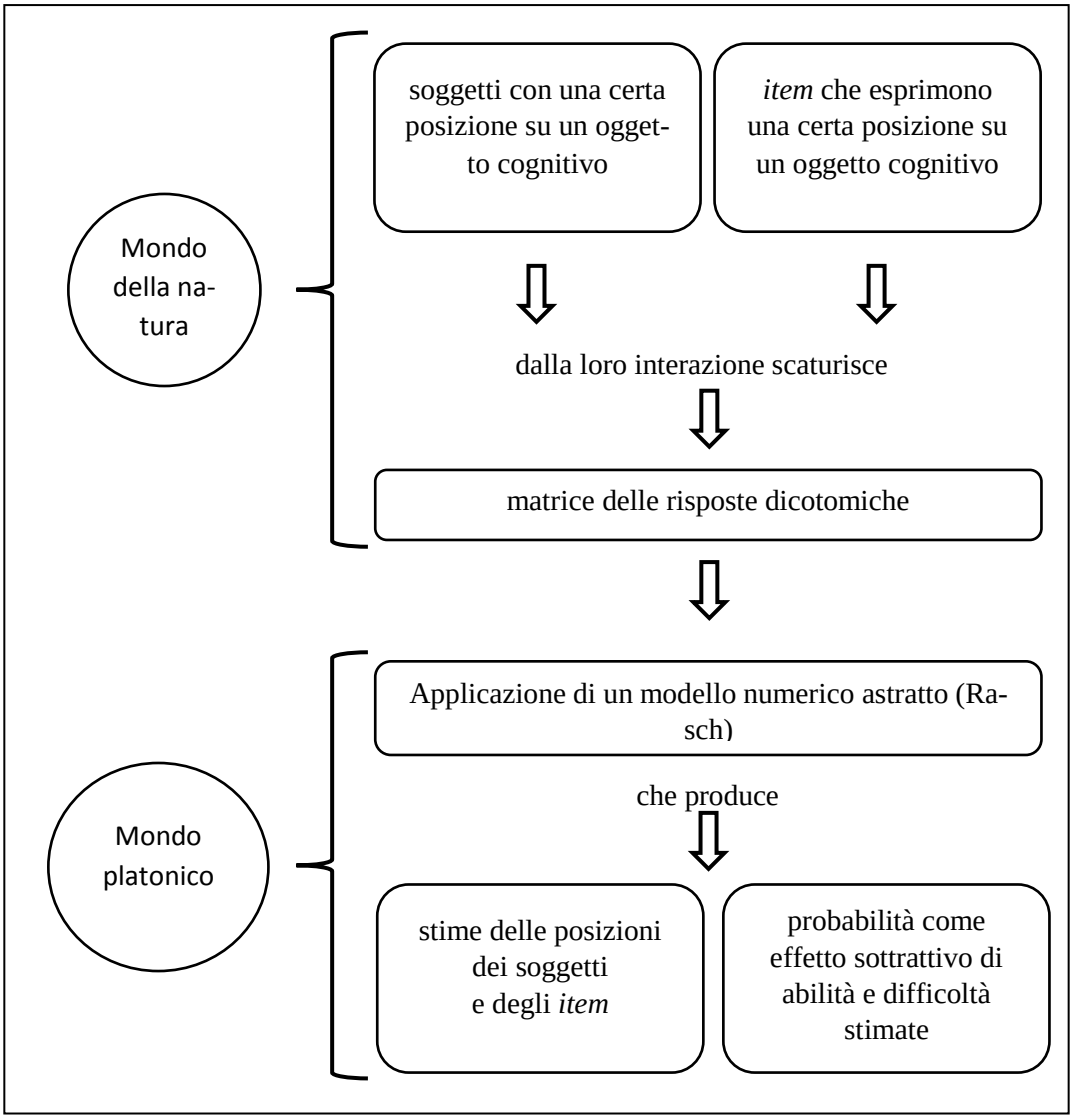


Fig. 12.2 – Nostra interpretazione del passaggio dal "mondo della natura" al "mondo platonico" nelle procedure coinvolte dal modello di Rasch, a partire dall'argomento di Kyngdon

In conformità a questo ragionamento, Kyngdon mostra perché, dal suo punto di vista, essendo gli assiomi della misurazione congiunta applicati a una matrice composta di stime e probabilità, ossia insiemi numerici, a rigore non staremmo testando le relazioni d'ordine su una struttura empirica e quindi verremmo meno a uno dei capisaldi della teoria ACM.

A quanto detto possiamo aggiungere un altro aspetto, ispirato dal tipo di argomenti di Kyngdon. Per introdurlo è necessario riconnettersi alla discussione precedente sull'aderenza del modello di Rasch agli assiomi. La funzione logistica che mette in relazione le probabilità congiuntamente alle stime dei parametri di abilità e difficoltà, ri-

spetta sempre gli assiomi della ACM e per questo è un modello *matematico* di misurazione congiunta[123]. Allo stesso tempo, il modello di Rasch è anche lo strumento attraverso il quale si trasformano le relazioni empiriche osservate di "superamento / non superamento di un *item*" e si rappresentano numericamente. Lo stesso strumento gioca quindi diversi ruoli in commedia: descrive la funzione del teorema di rappresentazione ed è la base per il test degli assiomi che, per la sua natura formale, non può che soddisfare.

Per cui, riassumendo, è *necessario* che il modello di Rasch produca delle stime (nel senso che le produce sempre) su qualsiasi base empirica; una volta prodotte le stime, è *necessario* che esse rispettino gli assiomi, a causa della struttura stessa del modello; una volta testati quelli, è *necessario* concludere che il modello descrive una struttura di misurazione congiunta additiva. Il ragionamento finisce per avvitarsi in modo circolare e tautologico e, a partire dalle stime, avviene su un piano matematico e logico di tipo formale e astratto: nella matrice non ci sarebbero oggetti che stiamo confrontando tra loro, ma numeri frutto di un modello matematico. La natura tautologica dell'argomento con cui si dichiara che il modello di Rasch è una forma di misurazione congiunta porterebbe a una conseguenza diretta di grande impatto sul piano dell'argomentazione scientifica: come caso formale di modello ACM, non esiste una fattispecie possibile di falsificazione degli assiomi perché questi ultimi, che dovrebbero rappresentare il criterio di controllo esterno, sono assorbiti dal modello stesso, interni a esso. Il possibile criterio di falsificazione slitta su altri aspetti e, non a caso, si attesta sul versante dell'adattamento dei dati al modello, che rimane la vera pietra di paragone tra sfera empirica e rappresentazione formale. Lo scoglio, qui, è dato dall'alone d'incertezza che avvolge l'analisi del *fit* statistico [Karabatsos 2001; si veda la discussione sul punto in precedenza] e dal rischio di una sua costante sovrastima.

[123] Nel dibattito tra Kyngdon [2008a; 2008b] e Borsboom e Zand Scholten [2008], questi ultimi tengono a sottolineare che nemmeno il primo mette in discussione il fatto che *formalmente* il modello di Rasch assicuri sempre la coerenza rispetto agli assiomi ACM.

12.3 Un'ipotesi di corretta applicazione degli assiomi ACM

Kyngdon [2011] accusa la mancanza di teorie non numeriche della casualità, ossia la non definitiva chiarezza su come le probabilità possano rappresentare le strutture sottostanti agli eventi casuali. Siccome, sostiene, non sappiamo quale struttura empirica rappresentino le probabilità, non possiamo accertare la natura quantitativa di un attributo utilizzando direttamente il modello di Rasch. L'equazione del modello di Rasch, per esempio, sebbene possieda caratteristiche particolari come l'invarianza dei parametri e la sufficienza delle statistiche[124], non può essere testata direttamente perché non possiamo somministrare ripetutamente e indipendentemente lo stesso *item* alla stessa persona per studiarne la distribuzione di frequenza delle risposte.

Da un punto di vista rappresentazionale, secondo Kyngdon, è comunque possibile utilizzare gli assiomi dell'ACM per testare la natura quantitativa degli attributi, utilizzando le proporzioni di risposte dei soggetti agli *item* in modelli non parametrici[125]. Il passaggio fondamentale per rendere legittima quest'applicazione richiede di abbandonare il concetto di stima della *propensione individuale* verso un certo oggetto cognitivo, per abbracciare un'interpretazione che veda la proporzione di risposte corrette a un *item* come indicatore della *probabilità che un gruppo di soggetti* con una certa abilità superi quell'*item* [126]. In effetti, i punteggi grezzi sono stimatori ordinali non distorti dell'abilità/difficoltà: in questo quadro, le proporzioni di risposta non sono utilizzate per stimare parametri di persone e *item*, ma per controllare se abilità e difficoltà possono sostenere una rappresentazione scalare a intervalli. Per usare le parole di Kyngdon [2011, 484],

[124] Kyngdon, tra l'altro, sottolinea che i punteggi nel modello di Rasch sono statistiche sufficienti solo in caso di perfetto adattamento, per cui alla fine l'invarianza dipenderebbe dal test e dal campione, mentre in fisica non c'è necessariamente questa dipendenza dallo strumento (oggettività). Inoltre aggiunge che l'invarianza dei parametri non comporta necessariamente una struttura quantitativa, perché ciò vale anche per scale che rilevano attributi in modo ordinale.

[125] Esempi di modelli non parametrici sono sviluppati e presentati in Scheiblechner [1999] e Karabatsos [2001].

[126] In questo caso, di fatto Kyngdon abbraccia una delle concezioni tradizionali del rapporto tra eventi e probabilità: quella *frequentista*, per cui la probabilità di un evento è definita come la proporzione di successi dell'evento stesso quando il limite del numero di prove ripetute tende a infinito. Generalmente si riconoscono almeno altre due concezioni fondamentali: quella *classica* e quella *soggettivista* [cfr. Piccolo 2004].

> usare questi risultati formali sui dati di un test reale può apparire arduo, dal momento che le probabilità di risposta a un *item* sono entità numeriche. Non sono gli assiomi della misurazione congiunta limitati a oggetti ed eventi qualitativi? Le strutture empiriche possono essere basate su entità numeriche a condizione che queste siano state ottenute empiricamente [Luce e Narens, 1994]. Una percentuale di risposte corrette a un *item* di un campione di persone con la stessa abilità può essere considerata una stima empirica campionaria della probabilità di risposta per quell'*item*. Poiché le proporzioni sono state ottenute empiricamente, gli assiomi della misurazione congiunta possono essere applicati alle matrici di tali proporzioni.

Kyngdon, dunque, accetta la possibilità di sfruttare le proporzioni empiriche di risposte corrette per testare gli assiomi della ACM, ma al di fuori dei modelli parametrici della IRT e propone una modifica nell'interpretazione dei dati che li trasformi in indicatori empirici direttamente testabili, conservando quindi la coerenza del quadro teorico rappresentazionale.

D'altro canto è proprio il carattere stocastico e parametrico del modello di Rasch a rappresentare un elemento di forza nell'analisi di dati che praticamente mai possono rispettare i requisiti assiomatici di teorie deterministiche. Gli argomenti di Kyngdon, tuttavia, servono a notare che, da un punto di vista rigorosamente rappresentazionale, l'interpretazione del modello di Rasch in chiave di ACM potrebbe comportare delle aporie teoriche non presenti, invece, nell'interpretazione corrente della Teoria del tratto latente. Rimane aperta la questione di quanto in profondità la critica di Kyngdon possa colpire l'ipotesi che il modello di Rasch sia una fattispecie di ACM. Una riflessione su questo punto coinvolgerà le prossime due sezioni.

12.4 Un rappresentazionalismo "rigoroso" *vs*. uno "debole"?

Borsboom e Zand Scholten [2008] e Zand Scholten [2011] criticano Kyngdon rispetto alla questione della natura delle probabilità e delle stime, invitando ad abbraccia-

re piuttosto una certa dose di pragmatismo, che induca a guardare cosa *sta dietro* quei numeri più che i numeri in se stessi. Il concetto di probabilità, sostengono, è senz'altro complesso e ha ricevuto molte diverse interpretazioni, alcune delle quali lo vedono come qualcosa di diverso da semplici "insiemi puri del reame platonico". In particolare, condividono l'interpretazione della probabilità in chiave di *propensione* [Popper 1959; Hacking 1965]: le probabilità sarebbero individuabili in entità reali, propensioni a fare o non fare, operanti nella realtà, capaci di sostenere la rappresentazione su un sistema relazionale numerico, sempre che possano essere percepite le loro relazioni. Questo, secondo Borsboom e Zand Scholten, è ciò che ricercherebbero i procedimenti di stima e la concezione della probabilità in chiave di propensione sarebbe quella adottata dai ricercatori applicati che utilizzano modelli come quello di Rasch.

Questo li spinge a definire due diversi approcci nell'adesione alla teoria rappresentazionale, uno "rigoroso" e uno "debole", sulla base di ciò che si è disposti a definire come "evidente" a livello empirico:

> Per il rappresentazionalista rigoroso, 'evidente' significa 'direttamente osservabile'. Per questo rappresentazionalista è fuori questione che qualsiasi modello probabilistico possa 'essere' misurazione congiunta, poiché le relazioni probabilistiche non sono osservabili direttamente (…). Per il rappresentazionalista debole, tuttavia, 'evidente' può significare 'evidente in via di principio', o 'evidente per un osservatore ideale', o forse perfino 'vero'. Un tale rappresentazionalista debole può forse supporre che le persone e gli *item* siano effettivamente ordinati dalle probabilità, e catturare questa ipotesi in un modello a variabile latente. (…). Siamo d'accordo che questo può non essere nello 'spirito' della RMT e che la posizione può condurre a serie difficoltà concettuali [Borsboom 2005] (…). La domanda è se un tale rappresentazionalista debole incontrerebbe difficoltà insormontabili nel fare ricerca empirica e costruire modelli di misurazione, o se potessimo indicare un'autentica incoerenza nel suo ragionamento. Ne dubitiamo. [Borsboom e Zand Scholten 2008, 115]

Il suggerimento avanzato è di concedere che probabilità e stime non siano solo insiemi puri numerici, bensì espressioni di realtà empiriche, rappresentazioni di fatti ed eventi e non solo astrazioni matematiche. Una diversa interpretazione degli elementi del modello eliminerebbe, almeno in parte, le incongruenze epistemologiche che incontra una visione rigorosa della RMT. Si tratta di un approccio diverso: un modello statistico formalizza una serie d'ipotesi su un qualche fenomeno, in modo testabile; se i dati empirici sono coerenti con il modello, cioè soddisfano le ipotesi iniziali, allora lo si può interpretare come una riproduzione delle relazioni che caratterizzano il fenomeno in questione. Diversamente, si deve concludere che una o più ipotesi o la loro congiunzione, sono smentite dai dati.

12.5 Evidenza empirica e strutture relazionali: un quadro complesso

La critica al modello di Rasch come forma di ACM portata da Kyngdon sul versante di un'interpretazione rigida della RMT, ci pare abbia molto a che fare con l'interpretazione della natura del dominio dei sistemi relazionali empirici coinvolti: le stime numeriche e le probabilità sarebbero numeri di cui non conosciamo il rapporto con la struttura empirica sottostante, anche se ci attestassimo sull'interpretazione della probabilità come *propensione* individuale (non potendo predisporre prove ripetute e indipendenti su uno stesso soggetto).

Ora, la questione della natura delle strutture empiriche è molto complessa e controversa e va al di là degli scopi di questo testo. Vale però la pena riflettere sul fatto che non è sempre così facile (come potrebbe a tratti sembrare dalla riflessione di Kyngdon) segnare un discrimine netto tra concetti che richiamano strutture empiriche e concetti che richiamano strutture formali astratte. Un richiamo a tale complessità è contenuto nel contributo di Sherry [2011] che affronta proprio il problema della definizione delle strutture empiriche, non solo nell'ambito della misurazione intensiva ma anche in quello della misurazione estensiva.

Il cuore della sua tesi risiede nella differenza sostanziale che si avrebbe tra la precisione con cui è possibile definire i due poli del processo di rappresentazione: gli elementi di un insieme numerico sono identificabili con una precisione molto alta, soprat-

tutto se, invece, comparata con la capacità di identificare e delimitare gli oggetti empirici. Sherry fa l'esempio delle procedure di misurazione del peso, ma il discorso può essere allargato a qualsiasi proprietà estensiva. Il dominio di un sistema relazionale empirico è un insieme indefinitamente grande di oggetti cui poter assegnare numeri sulla base della proprietà quantitativa da misurare. Tra questi ci sono anche casi-limite, che possiamo solo presumibilmente definire come misurabili, ma che empiricamente non lo sono: pesare una montagna, o misurare lunghezze oltre spazio conosciuto, per esempio, sono operazioni ammissibili in via di principio ma del tutto inattuabili, perlomeno nelle condizioni attuali. Ciò impedisce agli assiomi della misurazione di essere testabili empiricamente in modo esaustivo su un determinato dominio attraverso una generalizzazione di tipo induttivo. Inoltre, come si è già visto, anche non considerando l'aspetto dell'errore di misurazione, due degli assiomi della ACM (risolvibilità e condizione archimedea) pongono problemi di accertabilità nell'ambito d'insiemi finiti di elementi. Secondo Sherry, si procederebbe sempre secondo una logica di tipo abduttivo, in base a un principio di utilità pragmatica e spinti dai vantaggi che la misurabilità comporta in ambito scientifico: pur in assenza di un'inferenza induttiva degli assiomi su tutti i possibili oggetti, si adottano le procedure della misurazione, in virtù dei benefici che l'istituzione di una relazione tra i numeri reali e le strutture empiriche porta all'impresa scientifica, per quanto la struttura empirica non sia mai perfettamente corrispondente a quella numerica.

Anche la misurazione intensiva seguirebbe questa logica. Il teorema di rappresentazione stabilirebbe una relazione tra quello che Sherry definisce un "sistema quasi-empirico" e uno numerico. Che cosa intende Sherry con il concetto di "sistema quasi-empirico"?

> Qui un sistema quasi-empirico è una controparte insiemistica di un dominio di investigazione empirica, costruito sostituendo dominî e concetti matematici esatti[127] a quelli empirici inesatti. Non può esserci alcun isomorfismo strutturale tra un sistema relazionale genuinamente

[127] Precisiamo che in questo contesto il termine "esatto" (traduzione dall'inglese *exact*) non s'intende nell'accezione, diffusa nel linguaggio comune, di "esente da errori" o addirittura "vero", bensì di "preciso, minuzioso, rigido". Anche il termine opposto, "inesatto", non significa "errato, falso" ma "non preciso, non minuzioso, non rigido".

> empirico – cioè, un sistema i cui concetti sono strettamente empirici – e un sistema relazionale numerico. Poiché i concetti empirici, che sono inesatti, obbediscono a una logica differente dai concetti matematici, che sono esatti (…) ammettono casi-limite (…). In verità, un teorema di rappresentazione descrive gli assunti idealizzati che noi imponiamo a un sistema relazionale empirico al fine di produrre un sistema ipotetico-deduttivo che si possa avvantaggiare del potere inferenziale dell'aritmetica. [Sherry 2011, 520]

In sostanza, secondo Sherry senza "purificare" i concetti del dominio empirico, cioè senza trasformarli nella direzione della precisione e della rigidità di quelli del dominio numerico, non è possibile costruire una rappresentazione, un ponte tra le sfere empirica e numerica. Krantz, Luce, Suppes e Tverski [1971] riconoscevano l'eventualità di una qualche idealizzazione delle relazioni tra gli elementi di un dominio empirico, ma senza (sempre secondo Sherry [2011, 520]) porre l'accento debitamente che tale eventualità "sorge dall'emulare il sistema numerico stesso la cui struttura deve essere provata isomorfica al sistema relazionale empirico", cioè nasce come esigenza pragmatica di piegare la logica del mondo empirico a quella del mondo numerico. Le relazioni del dominio quasi-empirico non risiedono sulla superficie dei fenomeni ma derivano da una serie di operazioni concettuali preliminari.

Sherry giunge quindi a distinguere tra le affermazioni relative alla misurazione i) che riguardano predicati empirici, inesatti e che egli chiama "descrizioni" e ii) le affermazioni che riguardano predicati quasi-empirici, esatti, le quali sono "rappresentazioni". Come egli cerca di argomentare, solo le seconde sarebbero coinvolte nelle spiegazioni e nelle previsioni[128]. Alla fine sarebbero quindi considerazioni pragmatiche[129] e non osservative che farebbero invocare un certo livello di misurazione.

[128] Il corollario alla posizione di Sherry è che non sono le relazioni empiriche di superficie a poter falsificare una teoria, perché non è a quel livello, ma a quello quasi-empirico, che avvengono le spiegazioni.

[129] Incarnate dal processo di costruzione di un sistema quasi-empirico che rende *per definizione* i concetti empirici comparabili con quelli numerici.

Ricollegandoci all'argomento di Kyngdon e chiudendo la riflessione su esso, abbiamo rilevato che alla base del suo rifiuto dell'ipotesi che il modello di Rasch sia un caso di ACM si trova una definizione rigida e univoca di "sistema relazionale *empirico*". La tesi di Sherry aiuta a valorizzare la problematicità e la complessità di una tale definizione. Se fosse corretta la posizione per cui un teorema di rappresentazione coinvolge comunque una forma "idealizzata" di un sistema empirico, una specie di mediazione tra relazioni inesatte di superficie e relazioni esatte numeriche, allora la critica di Kyngdon alla natura non immediatamente empirica degli elementi utilizzati dal modello di Rasch finirebbe per essere chiaramente indebolita: adottato in modo rigido, tale ragionamento non troverebbe applicazione, secondo Sherry, in nessuna fattispecie di misurazione. Ovviamente, l'affermazione di quest'ultimo che in nessun caso i modelli di misurazione coinvolgono le relazioni empiriche di superficie è una critica anche alla posizione realista empirica, sostenuta da autori come Michell. E anche da quest'ultimo versante sono state avanzate critiche all'idea che il modello di Rasch sia un caso di misurazione additiva congiunta. Di queste ci apprestiamo a dare conto nel successivo capitolo.

13 Modello di Rasch, misurazione congiunta e quantità misurabili: una critica da un punto di vista realista

13.1 Misurabilità, mandati delle scienze quantitative e "pitagorismo"

La concezione della misurazione nell'approccio realista empirico di Joel Michell è vicina a quella classica adottata nell'alveo delle scienze fisiche, come si è già avuto modo di illustrare all'inizio del primo capitolo, quando si è incontrato per la prima volta l'autore nei panni di sostenitore e commentatore di questa visione. In generale, volendo semplificare al massimo, si può dire che la differenza cruciale e paradigmatica tra la teoria rappresentazionale difesa da Kyngdon e il realismo empirico riguarda il ruolo dei *numeri*. Infatti, per Michell i numeri non sono elementi astratti, esterni allo spazio e al tempo, entità insiemistiche pure senza contenuto empirico, bensì sono parte della realtà stessa delle cose [Sherry 2011]: i numeri si trovano *in re*, sotto forma di *rapporti tra grandezze* della stessa quantità. Come chiarisce Michell [2008b, 120]

> poiché la teoria rappresentazionale richiede che i numeri entrino nella misurazione attraverso una mappatura tra sistemi naturali, da una parte, e strutture numeriche logicamente indipendenti, dall'altra, essa richiede che i numeri siano logicamente distinti da quei sistemi naturali. Il problema è che nel paradigma della misurazione (ossia la fisica) i relativi sistemi naturali (cioè gli attributi della scienza fisica, come la distanza, la massa e il tempo) esemplificano già i numeri reali come rapporti di grandezze. Quindi, i numeri sono intrinseci ai sistemi naturali misurati e non, come sostiene la teoria rappresentazionale, estrinseci.

La misurazione è una descrizione di tali rapporti e, non a caso, nel linguaggio realista empirico a essa ci si riferisce come al tentativo di *scoprire* e stimare i rapporti tra le grandezze [Michell 1997]: inutile far notare che si può "scoprire" solo qualcosa che già esiste. Nelle parole di Michell [*ibidem*, 356] troviamo riassunto il suo punto di vista:

> Nella scienza quantitativa gli attributi (come la velocità, la temperatura, la lunghezza ecc.) sono assunti come misurabili. Cioè, si teorizza che un attributo, come la lunghezza, abbia un tipo distintivo di struttura interna, cioè una struttura quantitativa. Gli attributi che hanno questo tipo di struttura sono chiamati quantità. Seguendo un uso consolidato, le istanze specifiche di una quantità sono chiamate grandezze di quella quantità (es. la lunghezza di questa pagina è una grandezza della quantità lunghezza). Le grandezze di una quantità sono misurabili perché, in virtù della struttura quantitativa, si pongono in relazioni (rapporti) l'una con l'altra che possono essere espresse come numeri reali.

In pratica, se le varie realizzazioni di un certo attributo possono essere rapportate tra loro ottenendo dei numeri reali (cioè, se possono essere misurate), allora esse sono grandezze dell'attributo, il quale a sua volta è quantitativo. I numeri reali, così intesi, si troverebbero già nella struttura dell'attributo quantitativo, indipendentemente dal fatto che un misuratore esterno si disponga o no a rilevare il rapporto tra due grandezze specifiche[130]. Da questo si evince anche un altro aspetto fondamentale: che un attributo sia quantitativo e misurabile sarebbe un fatto *testabile*, cioè un'ipotesi da sottoporre a controllo empirico, che può dare esito positivo o negativo. E, ancora, la misurazione come stima del rapporto di una grandezza di un attributo quantitativo rispetto a un'altra dello stesso attributo (definita come *unità di misura*) non sarebbe un concetto scientifico a-teorico ma richiederebbe, al contrario, una teoria sottostante che riguardi le più generali ipotesi *empiricamente testabili* sulle relazioni tra attributi [Michell 1997].

Orbene, secondo Michell una scienza quantitativa ha due mandati[131] (*tasks*) cui assolvere: il mandato *scientifico* e il mandato *strumentale*. Il mandato scientifico riguarda il compito di investigare e accertare empiricamente la natura quantitativa degli attributi;

[130] Questo snodo è fondamentale per comprendere una differenza profonda tra il realismo *à la* Michell e l'operazionismo abbracciato da Stevens: per il primo l'oggetto della misurazione è *indipendente* dall'osservatore, per il secondo l'oggetto è definito dalle *procedure* del ricercatore.

[131] Si utilizza qui la traduzione "mandato" per l'inglese "*task*", piuttosto che un termine come "compito" o "funzione", per sottolineare il carattere di marcata obbligatorietà che questi aspetti rivestono nel quadro teorico di Michell.

il mandato strumentale concerne invece il compito di elaborare e costruire, una volta assodata la natura quantitativa di un attributo, le procedure e gli strumenti per misurare le sue grandezze [Michell 1997; 2000].

Gli attributi psicologici non sono direttamente definibili su un oggetto e manipolabili come quelli estensivi (es. la lunghezza o il peso): ciò che osserviamo sono le risposte a certi *item*. Nella pratica diffusa di ricerca, sostiene Michell, parallelamente a ciò si suppone che gli attributi teorici influenzino le risposte e s'ipotizza che la relazione tra punteggi e attributi sia una relazione tra quantità. La critica che l'autore avanza alla pratica delle scienze psico-sociali è definita, come abbiamo già scritto nel capitolo 8, dal concetto di "scienza patologica". A questo punto possiamo inserirlo in un quadro più generale. Infatti, il motivo per cui le scienze che ruotano attorno alle tecniche psicometriche sarebbero "patologiche" è, in definitiva, il tradimento del loro mandato scientifico: verrebbero meno al dovere di dimostrare l'ipotesi empirica che gli attributi che studiano siano quantitativi, il che solo potrebbe giustificare l'utilizzo legittimo dei modelli di misurazione, la cui costruzione ricade nel mandato strumentale.

Tale "tradizione" di rimozione avrebbe le sue radici in un'impostazione che Michell definisce "pitagorismo"[132], introdotta in ambito psicometrico dai pionieri della disciplina, in particolare Fechner. La *tesi della misurabilità* (gli attributi psicologici, o almeno alcuni di essi, sono misurabili) avrebbe sconfitto a tavolino l'*obiezione della quantità* (per cui nessun attributo psicologico è quantitativo e misurabile) [Sherry 2011] perché sarebbe stata accettata fin da subito la tesi "pitagorica" che *tutto* sia quantità e che quindi ogni cosa sia misurabile purché si individui la procedura per trasformare le osservazioni in numeri. Questo s'intende come effetto del pitagorismo: il rifiuto a priori dell'obiezione della quantità e l'idea, logicamente conseguente, che non esisterebbe in ultima istanza nessun mandato scientifico (nel lessico michelliano) ma solo uno strumentale volto a costruire modelli numerici per misurare gli attributi psicologici [Michell 1997]. Lo stesso accadrebbe con la definizione di misurazione di Stevens, che si concentra sulla regola di assegnazione dei numeri agli oggetti, ignorando il mandato scientifico di chiarire prima la natura degli attributi. Allo stesso modo, se un modello si adat-

[132] Da Pitagora, il grande filosofo e matematico greco del VI sec. a.C., il motto della cui scuola si dice fosse "Tutto è Numero", intendendo il numero come costitutivo della realtà stessa e dell'armonia di tutte le cose.

ta ai dati, questi sarebbero automaticamente considerati come frutto di attributi quantitativi, ignorando che quest'ultima è un'ipotesi che in quel modo non è ancora stata per niente testata.

13.2 Requisiti di un attributo quantitativo

Perché un attributo sia quantitativo è quindi necessario, per Michell, che possegga una certa struttura. Già all'inizio del Novecento, "Hölder rese esplicito come la struttura quantitativa coinvolga l'additività e che l'essere additivo di un attributo è una specifica condizione empirica" [Michell 2000, 649]. Hölder stesso illustrò la serie di assiomi necessari a definire quantità continue non limitate. Nonostante l'approccio filosofico alternativo su cui fonda i concetti di quantità e misurazione, Michell [2008b] riconosce pienamente l'importanza del contenuto matematico sviluppato nell'ambito della RMT. In particolare, la teoria ACM, se slegata dalla visione rappresentazionale dei numeri, svolgerebbe un ruolo "rivoluzionario" perché offre una base molto generale per dimostrare la natura quantitativa e additiva delle grandezze, comprese quelle derivate in fisica e, ciò che interessa al nostro discorso, gli attributi psicologici intensivi.

Il controllo degli assunti della misurazione additiva congiunta è, secondo Michell, un modo per sottoporre a test empirico gli attributi psicologici e risolvere la questione della loro natura. I modelli della *Item Response Theory* non risolverebbero di per sé la condizione patologica delle scienze psico-sociali. Infatti, il carattere continuo e quantitativo degli attributi trattati è assunto nell'utilizzo della funzione caratteristica dell'*item* per descrivere il rapporto presunto tra attributo e probabilità, senza che esso sia prima provato dal punto di vista empirico; inoltre, questo assunto non è spesso nemmeno esplicitato e formalizzato in modo aperto ma dato per scontato[133].

[133] Dello stesso avviso anche Heene [2013] che, oltre a concordare sul fatto che la ricerca *mainstream* in ambito IRT assumere la natura quantitativa delle dimensioni latenti senza testarla seriamente e rimuovendo il tema dal quadro teorico, rileva anche una forte tendenza a evitare la falsificazione dei modelli: soprattutto in campo psicologico e psichiatrico sarebbero molto rari gli articoli che riportano una falsificazione delle ipotesi. Ciò sarebbe, secondo l'autore, ancora più evidente nel caso della IRT, dove al posto di controlli stringenti sui modelli si utilizzano i meno efficaci test di adattamento. Esempi di meta-analisi che mettono a confronto i risultati di ricerca nelle pubblicazioni in vari ambiti scientifici si trovano in Fanelli [2009; 2010].

Michell [2008b] nota tuttavia che, a differenza dei modelli IRT più articolati che utilizzano un numero maggiore di parametri per aumentare l'adattamento del modello ai dati, quello di Rasch sarebbe utilizzato dai suoi sostenitori soprattutto come criterio per costruire scale con l'ambizione di produrre misurazioni. Da una parte, quindi, i modelli a due e tre parametri sarebbero più che altro rivolti ad *adattarsi ai dati* per riprodurre una determinata situazione nei termini degli attributi e dei parametri coinvolti; il modello di Rasch sarebbe invece più utile per stabilire *quali dati si adattano al modello* e selezionare gli *item* che possono costituire scale a intervalli. Ciò non toglie, come vedremo tra poco, che l'adattamento al modello di Rasch non può essere, secondo Michell, la garanzia che la struttura dell'attributo rilevato sia quantitativa, perché la natura degli attributi è da dimostrare indipendentemente, al di fuori e al di là del modello.

In generale, in polemica con la posizione di chi, come Borsboom e Mellenbergh [2004], appoggia la tesi della superiorità dei modelli probabilistici rispetto a quelli deterministici, Michell rileva che in realtà se un insieme di dati si adatta ai secondi, generalmente, lo fa anche con i primi: non si tratta quindi di stabilire *a priori* quali modelli siano più potenziali di altri, bensì di sviluppare teorie sugli attributi che ne giustifichino la dimensionalità empirica, così da rendere la questione della scelta dello strumento di analisi una di *appropriatezza* e non di superiorità in astratto. Questo è quello che, secondo l'autore, non avverrebbe nelle scienze psico-sociali. Michell [2004] esclude quindi che l'analisi del *fit* statistico possa essere, se ce n'è uno, lo strumento conclusivo attraverso il quale stabilire la natura quantitativa degli attributi: pur se un modello si adattasse perfettamente ai dati, ciò non potrebbe garantire di per sé la natura quantitativa di un tratto latente[134].

[134] Vautier *et al.* [2012] sostengono, utilizzando il concetto di falsificazione di Popper, che nelle teorie psicometriche si trovano solitamente proposizioni che non possono essere falsificate perché si adattano a ogni circostanza, cioè saturano in ogni caso il sistema osservativo di riferimento. Pensando a un modello probabilistico, in effetti, non ci sono casi empirici che esso consideri impossibili perché la probabilità non è mai esattamente '0' o esattamente '1' per alcun risultato e quindi ogni caso ha sempre una qualche probabilità di compiersi. L'unico modo di introdurre un principio di falsificazione sarebbe stabilire un limite oltre il quale l'adattamento dei dati è ritenuto non accettabile, ma per statuire una soglia precisa sarebbe necessaria un'altrettanto precisa motivazione teorica per collocarla proprio in quel punto e non in un altro.

13.3 Verifica della natura quantitativa dell'attributo: inclusione ed esclusione degli *item* e analisi dell'adattamento

In prima istanza, ciò che potrebbe impedire di corroborare la natura quantitativa della proprietà, anche alla presenza di un *fit* molto alto, è l'eventuale debolezza delle teorie sugli attributi. Michell [2004, 124] individua la condizione ideale nella ricerca sperimentale, dove "la relativa teoria deve essere sufficientemente dettagliata per specificare le varietà di situazioni riguardo alle quali possono essere dedotte le previsioni". Nell'ambito dei modelli psicometrici, in buona parte tale ruolo di specificazione spetta alla selezione degli *item*; i modelli in sé, come costruzioni matematiche, niente avanzano sulla natura e sul carattere qualitativo degli attributi, fatti salvi i requisiti della continuità e della quantità. Sono gli *item* che, organizzati nel test attitudinale, rappresentano una fattispecie delle "situazioni" che caratterizzano l'attributo e rispetto alle quali produrre le previsioni. Ma, per l'appunto, la debolezza attuale della generalità delle teorie sugli attributi[135] farebbe sì di non andare oltre l'elaborazione e la selezione degli *item* secondo criteri informali, cioè senza un rapporto di consequenzialità stringente e diretta tra teoria e test[136]. Inoltre, come già discusso in precedenza, è pratica comune quella di eliminare dal test (non in fase di costruzione dello strumento, ma in fase di analisi dei dati) gli *item* che tendono a deprimere il valore di adattamento complessivo ai dati.

Ora, per quanto riguarda la problematica dei processi di costruzione degli *item*, la posizione di Michell appare forse troppo rigida, poiché se è vero che la discrezionalità del ricercatore gioca in questa fase un ruolo importante, è vero anche che le scelte pos-

[135] Kyngdon stesso [2011] sostiene la necessità di accompagnare la verifica degli assiomi ACM con teorie descrittive che, però, al momento sarebbero molto carenti in ambito psicometrico.

[136] A conclusioni simili, anche se di portata più ampia, giunge anche Sijtsma [2012, 787], affermando che "una misurazione significativa è possibile solo se si conosce a sufficienza l'attributo così da giustificare la sua operazionalizzazione logica in prescrizioni dalle quali possa essere sviluppato uno strumento di misurazione. Un problema immenso in psicologia è che le teorie sugli attributi sono spesso non abbastanza precise da giustificare un'operazionalizzazione logica. Entrambe le prospettive fisicista e statistica [la prima è quella della teoria ACM, la seconda quella dei modelli probabilistici della IRT, N.d.A.] pongono poca attenzione a questo problema. La prospettiva fisicista assume che le teorie psicologiche sugli attributi possano raggiungere un alto grado di precisione comparabile a quello delle teorie sugli attributi fisici, ma al momento questo è un obiettivo inattuabile. La prospettiva statistica assume che gli attributi abbiano la struttura dei modelli IRT, ma ignora che questo assunto non è basato su teorie sostanziali ben sviluppate su attributi specifici e che, eccetto casi rari, non c'è nessuna prova convincente per la congruenza assunta."

sono (e dovrebbero) sempre essere esplicitate, giustificate, discusse alla luce della definizione dell'attributo, per quanto le procedure possano non essere strettamente formali nel senso delle scienze sperimentali: l'alternativa alla grande capacità di specificazione della fisica e a criteri rigidi ed esatti di selezione rigorosamente formalizzati, non è necessariamente l'elaborazione casuale, scriteriata e del tutto informale, senza alcuna possibilità di controllo.

Per quanto riguarda invece il tema dell'eliminazione degli *item* del test per aumentarne il *fit*, si è già osservato che modificare il quadro delle affermazioni che rappresentano la scala dell'attributo richiede anche necessariamente una ridefinizione più o meno rilevante dell'attributo stesso e della sua struttura semantica: quando ciò non accade, si corre il rischio di essere fuorviati e si può imputare un carattere quantitativo all'attributo quando si è alla presenza di uno al massimo ordinale[137]. Poiché si è mostrato [Michell 1994] che dati la cui struttura è semplicemente ordinale possono adattarsi perfettamente ai rigidi ed esigenti modelli di tipo deterministico, si può facilmente immaginare che questo potrebbe accadere a maggior ragione con i modelli probabilistici, che hanno requisiti più blandi dei primi.

13.4 Verifica della natura quantitativa dell'attributo: le conseguenze attese

Questo ci porta al secondo aspetto sollevato da Michell riguardo alla verifica della natura quantitativa di un attributo. Se da un lato è necessaria una teoria in virtù della quale selezionare in modo rigoroso e coerente gli *item* del test, dall'altro è altresì necessario possedere dei criteri per discernere le conseguenze che contraddistinguono una struttura quantitativa (cosa necessariamente ci aspettiamo per sancire una struttura genuinamente additiva) rispetto a una semplicemente ordinale.

L'ACM corre in soccorso proprio per stabilire questi aspetti e, come si è discusso, la gerarchia delle cancellazioni (in particolare la cancellazione doppia) è il cuore pulsan-

[137] Scrive, infatti, il nostro autore che "quand'anche l'attributo psicologico sottostante fosse niente più che un ordine parziale, potrebbe essere selezionato un sottoinsieme strettamente ordinato di *item*, e i modelli di risposta all'*item* potrebbero adattarsi bene ai dati conseguenti. Cioè, questi modelli possono adattarsi anche quando gli attributi relativi sono non quantitativi." [Michell 2004, 125] Ciò significa che selezionando un sottoinsieme di *item*, magari *ad hoc*, si potrebbe essere in grado, in alcuni casi, di trovare quelli che si adattano molto bene al modello di Rasch, anche se riferiti a una dimensione non quantitativa.

te della questione. Ricordiamo di nuovo le parole di Michell [2008b, 17], che mette in luce l'esistenza di due gruppi di relazioni tra le coppie ordinate di oggetti, uno solo dei quali distingue il caso in cui "la relazione d'ordine tra le coppie di rapporti è determinata dalla struttura delle grandezze al di là del semplice ordine, ciò che potremmo chiamare la *struttura additiva dell'attributo*", mentre l'altro riguarda solamente il caso in cui "la relazione d'ordine tra le coppie di rapporti è determinata dall'ordine delle grandezze coinvolte". Ordine e quantità sono strutture diverse la cui natura emerge direttamente dall'indagine empirica degli assiomi dell'additività, che nel caso di grandezze non concatenabili ricadono nell'ambito d'ispezione dell'ACM.

Da un punto di vista squisitamente matematico, il modello di Rasch, come fattispecie a un solo parametro di modello di risposta all'*item*, è in grado di discernere tra strutture ordinali e quantitative perché soddisfa la necessaria gerarchia di cancellazioni. Inoltre, il campo del modello di Rasch e le relazioni tra i suoi elementi (la probabilità di superare un *item* come prodotto non interattivo di abilità e difficoltà) sono esattamente un caso cui poter applicare in modo appropriato gli assiomi dell'ACM. Eppure, secondo Michell [2008b] vi è una netta differenza tra quest'ultima e il modello di Rasch, che rende inappropriata l'identificazione tra i due.

Innanzitutto, sostiene, se è vero che la relazione tra probabilità, abilità e difficoltà è un caso di specie la cui testabilità ricade nell'ambito di genere della misurazione additiva congiunta, è altrettanto vero che la teoria ACM ha una portata più generale che riguarda tutti i possibili casi in cui si abbiano tre attributi, di cui uno in funzione non interattiva degli altri. Questo primo aspetto individuato da Michell, per quanto possa essere rilevante, si sofferma sul contenuto inserito nei modelli (cioè la fonte dei dati che si utilizzano) e ci sembra molto meno cruciale rispetto al secondo punto, che affronta la questione della struttura formale su cui s'incardinano modello di Rasch e ACM. A questo proposito, infatti, Michell osserva una distinzione netta: gli assiomi della teoria della misurazione additiva congiunta controllano le *condizioni necessarie a livello di relazioni ordinali* affinché tre attributi mostrino una struttura quantitativa, indaga cioè l'esistenza o meno di quest'ultima nelle dimensioni in esame; il modello di Rasch, invece, ipotizza una certa relazione tra gli attributi *utilizzando una funzione matematica specifica e particolare*, giungendo alla gerarchia di cancellazioni proprio grazie al suo utilizzo e, infine, assumendo come precondizione (non verificata e spesso implicita) la

struttura continua delle dimensioni latenti, senza alcun reale controllo sulle relazioni d'ordine. Si tratta di due procedimenti diversi da un punto di vista logico e sostanziale che, nonostante le affinità di superficie, portano Michell a rifiutare la tesi che il modello di Rasch sia un esempio di ACM e, soprattutto, un modo equivalente di testarne gli assiomi.

13.5 Le pretese dei modelli probabilistici

E' qui che, a nostro parere, s'inserisce la critica più radicale di Michell. Il nodo del contendere può essere posto in questi termini: da cosa deriva la pretesa dei modelli probabilistici di produrre misurazioni su scale a intervalli e di essere una rappresentazione più fedele della realtà rispetto a modelli ordinali? Come abbiamo discusso in altri punti del testo, il modello di Guttman, per fare un esempio, è interpretabile come un caso-limite del modello di Rasch, ottenuto nella situazione in cui il coefficiente di discriminazione del modello è infinito. In questa fattispecie, il modello diventa deterministico (in caso di perfetta scalabilità, la precisione nello stabilire il livello di un soggetto è massima, avviene senza errore e dal punteggio si può ricostruire lo schema di risposte) e il livello di "misurazione" ordinale (è possibile stabilire l'ordine di affermazioni e soggetti, ma non le distanze tra loro). Che cosa differenzia quindi questo caso-limite dal modello di Rasch? Che cosa permette di passare dalle scale ordinali dell'uno a quelle a intervalli dell'altro? Secondo Michell [2004] non è una maggiore o minore conoscenza dell'attributo alla base del modello, bensì l'adozione dell'ipotetica forma funzionale che stabilisce un rapporto tra i livelli degli attributi e la probabilità di risposta. Questa forma funzionale è sostanzialmente determinata dalla teoria dell'errore casuale associata al modello[138]. Michell si domanda: cosa sappiamo realmente dell'errore e in che modo attraverso esso ci possiamo spingere a fare affermazioni sulla natura degli attributi? La sua risposta è che non conosceremmo niente dell'errore e che nessuna pretesa di scoprire la natura quantitativa della struttura degli attributi per questa via possa essere reclamata.

[138] Michell scrive, utilizzando una metafora a fini esplicativi, che "l'errore è la finestra attraverso la quale si pensa di contemplare la struttura quantitativa." [2008b, 122]

> Ora, se la risposta corretta di una persona a un *item* dipendesse unicamente dall'abilità, senza alcuna componente di 'errore' casuale coinvolta, uno potrebbe solamente apprendere il fatto ordinale che l'abilità di quella persona come minimo corrisponde al livello di difficoltà dell'*item* [caso del modello di Guttman, N.d.A.]. Gli utilizzatori di modelli di risposta all'*item* derivano tutta l'informazione quantitativa (come distinta da quella meramente ordinale) dalle proprietà della distribuzione della componente dell''errore' casuale[139]. Se il modello è vero, la forma della distribuzione dell''errore' riflette la struttura quantitativa dell'attributo, ma se l'attributo non è quantitativo, la forma supposta dell''errore' proietta solamente l'immagine di una quantità fittizia. Qui, come altrove, gli psicometrici derivano ciò che essi vogliono di più (le misure) da ciò che essi conoscono di meno (la forma dell''errore'), attraverso la presunzione di conoscerlo già. [Michell 2004, 126]

Quest'ultimo argomento, incentrato sulla critica alla pretesa che l'errore possa essere lo strumento della quantificazione di un attributo, è anche la premessa per quello che Michell definisce il "paradosso di Rasch". Di quest'ultimo ci occuperemo adesso, affrontandolo nell'ultimo capitolo.

[139] Michell trova ciò improprio poiché, ricordiamo, secondo quanto egli stesso sottolinea della teoria assiomatica della misurazione congiunta, dovrebbe essere la gerarchia delle cancellazioni a stabilire le relazioni che individuano semplici strutture ordinali e quelle che accertano una struttura quantitativa, al di là della mera struttura ordinale.

14 Il paradosso dei modelli di Guttman e Rasch

14.1 L'argomento del paradosso

Che il modello di Rasch possa essere considerato, da un punto di vista rigoroso, una generalizzazione del modello di Guttman non è accettato da tutti gli studiosi, come vedremo tra poco. Per il momento seguiamo la linea di ragionamento di Michell, adottando l'interpretazione più volte richiamata anche nel corso di questo testo.

Nel modello di Guttman, se la posizione di un soggetto sulla dimensione latente è almeno pari a quella espressa da una frase, allora si assume che il primo accetterà di sicuro la seconda; se invece il grado di proprietà del soggetto è inferiore, di sicuro la rifiuterà. Il modello di Rasch, sostiene Michell [2008b], è una versione *woolly* del modello deterministico, ossia una fattispecie più sfocata, vaga o addirittura confusa: essa allenta i requisiti del modello di Guttman, inserendo un fattore aleatorio. Formalmente, se X è il livello di proprietà del soggetto e Y quello dell'*item*, nel modello di Guttman il soggetto accetterà correttamente l'*item* se e solo se $X \geq Y$, mentre nel modello di Rasch lo schema prevede che il soggetto accetti in modo corretto l'*item* se $X + e \geq Y$, dove il numero reale e è il fattore aleatorio che concede la possibilità che un soggetto con abilità inferiore alla difficoltà ($X < Y$) possa superare l'*item* o che uno con abilità maggiore ($X > Y$) possa non superarlo.

Sottolineiamo, a questo punto, che definire questo fattore come 'errore' può essere fuorviante poiché, come riconosce anche Michell, si tratta di uno scostamento rispetto a un modello teorico ed è quindi esso stesso un concetto puramente teorico: da un punto di vista logico, non è il soggetto che "sbaglia" rispetto al modello, ma quest'ultimo che non riesce a interpretare la risposta del soggetto come riproduzione deterministica di una funzione di risposta. Per questo, a maggior ragione se ci spostiamo dall'ambito dei test di abilità a quello dei test attitudinali, ci sembra più neutra e adeguata l'etichetta di "fattore aleatorio", che non sconta la carica semantica negativa insita nel termine 'errore'[140] e permette di distinguere nettamente i veri e propri errori di misurazione. Nel te-

[140] Marradi [2002], per esempio, per motivi analoghi utilizza il termine 'residuo' invece di 'errore' al riguardo dei modelli di regressione.

sto, "fattore aleatorio" ed "errore casuale" saranno quindi utilizzati con lo stesso significato.

Secondo Michell, dunque, la differenza tra i due modelli risiede nel fattore aleatorio. Conseguentemente, a questo fattore deve anche essere fatta risalire la differenza tra livello ordinale di "misurazione" del modello deterministico e livello intervallare di quello probabilistico. Qui, alla luce della logica generale alla base della misurazione, risiederebbe il paradosso di Rasch: *com'è possibile che, introducendo un fattore d'incertezza, aumenti il livello di precisione della misurazione*? Nelle scienze fisiche, diminuendo i fattori di aleatorietà fino ad annullarli (tendenzialmente), si avvicina asintoticamente una misura perfetta. Nei modelli psicometrici sembrerebbe invece avvenire il contrario:

> Nel modello di Rasch, se fossero eliminati gli errori, la misurazione quantitativa diventerebbe impossibile; se le condizioni delle prove potessero essere migliorate fino al punto dove $e = 0$ sempre, le misure di abilità non sarebbero migliorate. Il modello di Rasch si ridurrebbe a quello di Guttman e le abilità potrebbero solo essere ordinate. Ma eliminare completamente l'errore deve (per definizione) migliorare le nostre osservazioni. Perciò, se osservazioni affinate potrebbero solamente mostrare che le abilità sono meramente ordinate, come potrebbero osservazioni meno affinate fornire da sole prove che esse sono più di ciò? [Michell 2008b, 122].

Anche Kyngdon [2008b] riprende il paradosso e concorda con Michell, sostenendo che "se è perfettamente osservata" la struttura triangolare del modello di Guttman "allora le abilità sono solamente ordinali". Ma

> dato che l'osservazione empirica non è perfetta, l'errore nasconderà la struttura. Per gli psicometrici questo errore significa introdurre la probabilità nel modello. Così facendo, la relazione ordinale tra l'abilità e la difficoltà dell'*item* deve essere sostituita con una additiva o altrimenti non ha senso. Questa relazione additiva esige che l'abilità e la difficoltà siano quantitative – il che è in contrasto con l'osservazione perfetta. [Kyngdon 2008b, 128]

Zand Scholten [2011], seppur da annoverare tra i critici del paradosso, riporta un esempio molto esplicativo del suo significato. Immaginiamo che alcuni astronomi posseggano alcuni indizi visivi che attorno a una stella, individuata a lunga distanza dalla Terra, ruoti un sistema planetario, ma che l'osservazione sia disturbata da una nube di foschia. In seguito la nube scompare e l'osservazione non è più disturbata: gli astronomi costatano allora che il sistema planetario non è più visibile. Essi sono allora portati a rifiutare l'ipotesi dell'esistenza del sistema planetario, poiché con la scomparsa del fattore aleatorio (l'errore, cioè la nube) anche il sistema è scomparso, lasciando supporre che fosse solo un'apparenza illusoria. Allo stesso modo, possiamo noi accettare che un attributo sia quantitativo perché il modello di Rasch sostiene una misurazione a livello d'intervalli, quando eliminando l'errore casuale l'unica informazione rimanente è quella ordinale?

L'argomento è molto affascinante e spinge ulteriormente a riflettere sul tipo di misurazione consentito sugli attributi psicologici. La correttezza del paradosso implicherebbe una pesante messa in discussione del fatto che il modello di Rasch possa produrre scale a intervalli, anche solo in via di principio.

14.2 Critiche al paradosso. Il modello di Guttman è davvero un caso-limite di quello di Rasch?

C'è chi, però, ne mette in discussione la fondatezza. Ciò avviene principalmente secondo due strade: la prima cerca di dimostrare che in realtà il modello di Guttman non potrebbe essere ricondotto semplicemente a una fattispecie deterministica del modello di Rasch, la seconda che non necessariamente le fonti di errore o i fattori aleatori sono destinati a ridurre la precisione della misurazione.

14.2.1 Un punteggio, due interpretazioni

Sulla prima strada si pone per esempio Sijtsma [2012], che ritiene non ci sia nessun paradosso di Guttman-Rasch ma solo un utilizzo diverso dei punteggi. L'autore parte da un presupposto generale: nessun insieme di dati raccolti può essere esaustivo ri-

spetto al controllo di modelli continui e coincidere perfettamente con essi, giacché i primi sono necessariamente discreti e finiti e da essi non sarebbe quindi possibile concludere che una scala non abbia determinate proprietà[141]. L'errore nell'argomento di Michell sarebbe far partire il ragionamento sui modelli dalla struttura dei dati, invece che dalle differenze nei due modelli stessi. In virtù di questo quadro concettuale, secondo Sijtsma la differenza tra i modelli di Guttman e di Rasch non ha niente a che vedere con l'introduzione o meno dell'errore casuale nella funzione di risposta. Piuttosto, osserva, mentre nel modello di Guttman il punteggio di soggetti e *item* è usato come semplice indicatore ordinale delle posizioni, nel modello di Rasch esso è la statistica sufficiente per stimare i parametri su una scala a intervalli. Ciò sarebbe semplicemente la conseguenza di un diverso utilizzo delle stesse informazioni: il modello di Rasch considera significativa la differenza tra parametro del soggetto e quello dell'*item* e assume che sia in relazione monotona con la probabilità di superare il secondo, giungendo a un livello di misurazione a intervalli; il modello di Guttman, invece, rinuncerebbe a utilizzare questa informazione, il che giustifica il livello di misurazione più basso. Gli intervalli che utilizza il modello di Guttman possiedono quindi una maggiore autonomia strutturale e semantica tra loro e l'informazione, comunque racchiusa al loro interno, non è estratta dal modello. Si tratterebbe quindi di una logica di costi e benefici: un modello più complesso, che contempla il ruolo dell'errore casuale, ripaga con un livello di misurazione più elevato e viceversa.

Heene [2013] ribatte a Sijtsma che la sua tesi non attaccherebbe, in realtà, i cardini del paradosso di Guttman-Rasch. Infatti, argomenta, il suo ragionamento parte dal presupposto i) che l'attributo sia continuo e ii) che il modello di Guttman ignori una presunta informazione contenuta negli intervalli scanditi dai punteggi. Queste, però, sarebbero entrambe due ipotesi non dimostrate. Il punto i) è parte degli assunti dei modelli IRT e quindi andrebbe provato empiricamente come presupposto della misurazione intervallare. La non validità del punto ii) segue dal primo: non ha alcun senso parlare di informazioni contenute negli intervalli, che un modello può o no utilizzare, se l'esistenza di queste informazioni è solo ipotetica, cioè qualcosa che potrebbe non sussistere e che si accetta solo per *fiat*.

[141] Di un aspetto simile parla anche Michell nella citazione riportata in nota 137.

La questione rimane dunque irrisolta e connessa a doppio filo a ciò che gli studiosi sono o no disposti ad accettare come "assunto dimostrato" e con quale forza esso lo sia dalle conoscenze esistenti e dalla teoria.

14.2.2 Probabilità binaria e continua

Zand Scholten [2011] arriva a conclusioni simili a quelle di Sijtsma, accentuando la riflessione sul concetto di *continuità*. La differenza sostanziale tra il modello di Guttman e quello di Rasch non risiederebbe tanto nell'introduzione di un generico fattore aleatorio, bensì di uno di tipo continuo. Il modello di Guttman utilizza una forma di probabilità binaria, che contempla solamente una probabilità certa (*item* superato) e una nulla (*item* rifiutato), secondo una logica di tipo *discreto*. Ma questo tipo di approccio spegne necessariamente qualsiasi possibilità di distinguere la posizione di soggetti che hanno superato uno stesso *item* ma che potrebbero possedere livelli molto diversi di abilità[142]. Il modello di Rasch, invece, contempla tutto il *continuum* delle probabilità da 0 a 1, permettendo di stabilire una relazione diretta tra questo e la distanza tra abilità e difficoltà. In questo modello l'errore è di un tipo specifico, ossia dipende dal livello di abilità del soggetto. A riprova di ciò, vi sono alcuni modelli in cui i valori dell'errore sono costanti (o legati all'*item* o al valore atteso del punteggio) e che possono essere considerati come fattispecie di quello di Guttman cui si è aggiunto un errore: eppure, la loro natura costante non intacca il carattere discreto delle probabilità risultanti. Per cui, se da un punto di vista matematico si deve riconoscere che il modello di Guttman è un caso-limite di quello di Rasch rispetto alla discriminazione, per Zand Scholten non è però quella la differenza decisiva tra i due. Il carattere discreto *versus* quello continuo sarebbe il vero nodo della questione.

Questa conclusione ci pare del tutto condivisibile, ma non altrettanto la conseguenza che da ciò dovrebbe derivare, cioè che se non è l'errore in generale il cuore della questione e se Rasch non è semplicemente "Guttman più l'errore", allora l'argomento di Michell decadrebbe. Infatti, quest'ultimo affronta il tema dell'introduzione del fattore aleatorio necessariamente nella forma specifica in cui questo si presenta nel modello di

[142] Zand Scholten mostra con un esempio come la natura discreta delle probabilità utilizzate dal modello di Guttman fallisca nel soddisfare l'assioma di cancellazione doppia, a differenza invece del modello di Rasch in cui, per la sua forma funzionale, è sempre soddisfatto.

Rasch e non in altre. Una volta stabilito che esistono forme diverse di errore che possono essere applicate al modello di Guttman senza mutarne il carattere e che il vero "salto" che avviene rispetto a quello di Rasch riguarda l'aspetto continuo contro quello binario (discreto) delle probabilità, si deve però pur sempre costatare che tale "salto" avviene grazie all'introduzione di una funzione continua (logistica), giustificata dal presunto rapporto tra fattore aleatorio (errore casuale continuo) e differenza tra i parametri di abilità e difficoltà. Il fatto che non tutti i tipi di errore producano paradossi come quello di Guttman-Rasch, non significa che ciò non possa avvenire in quest'ultimo caso. Questo argomento di Zand Scholten introduce seri argomenti di riflessione e un punto di vista efficace sulla questione, ma non può dichiarare chiusa la questione rispetto al nodo, indicato da Michell, del rapporto tra attributi e fattore aleatorio.

14.3 Critiche al paradosso. Fattori aleatori e precisione: può l'errore migliorare la misurazione?

Ancor più interessante è un'altra via di ragionamento, che sempre Zand Scholten intraprende, sulla questione del fattore aleatorio e della precisione. La base su cui Michell porrebbe il paradosso di Rasch è, per l'autrice, ininfluente e di questo abbiamo appena parlato. Volendo, però, portare ad ampio raggio la sua critica al paradosso, accetta la sfida di porre, per ipotesi, che il fattore aleatorio sia una base efficace su cui fondarlo: è sempre vero che introdurre un elemento di "disturbo" diminuisca la precisione di una misurazione?

14.3.1 Quando l'errore aiuta la misura: un'analogia e una simulazione

L'autrice mostra alcuni esempi per supportare la tesi che, in via di principio, la risposta alla domanda precedente è negativa, perché vi sarebbero situazioni in cui l'introduzione di un "rumore", se dipendente dal tratto latente, può invece migliorare la misurazione. Lo spunto viene da campi esterni alle scienze sociali e in particolare dalla teoria della *risonanza stocastica* [es. McNamara, Wiesenfeld 1989; Simonotto et al. 1997; Gammaitoni et al. 1998]: in estrema sintesi, in determinate circostanze un segnale

ondulatorio debole può essere migliorato e reso rilevabile attraverso l'introduzione di un disturbo (*rumore*) che, entrando in risonanza, lo amplifica. Questo e altri esempi sono senz'altro elementi a supporto del fatto che una qualche forma di fattore aleatorio può servire ad aumentare la precisione della misurazione stessa, confutando quindi l'affermazione generale e di principio che ciò non possa mai accadere[143]; ma di per sé non sono certo risolutivi per dimostrare che lo stesso avvenga o possa avvenire anche in altri casi specifici: per dimostrare un'analogia del genere e "importare" un principio da un campo cognitivo a uno molto diverso si dovrebbe provare l'esistenza di una somiglianza reale tra i meccanismi coinvolti nei vari casi.

Zand Scholten [2011] fa un passo avanti e prova a mostrare che lo stesso si può osservare effettivamente in alcuni casi di misurazione sociale e, nello specifico, che ciò è matematicamente possibile anche nel modello di Rasch. Per farlo, utilizza un insieme di dati simulati, controllando l'andamento dell'errore quadratico medio del test in determinate situazioni. Innanzitutto costruisce un campione di 500 soggetti estratti da una popolazione fittizia distribuita normalmente rispetto all'abilità; immagina poi che essi rispondano, in accordo con la funzione del modello di Rasch, a 16 affermazioni di varia difficoltà. Come si è avuto modo di spiegare nella Parte Seconda, la pendenza della curva (governata dal parametro di discriminazione degli *item*) è direttamente connessa alla sensibilità con cui l'*item* riesce a distinguere l'abilità dei soggetti, in particolate nella regione di valori adiacente al suo livello di difficoltà: per questo, le prove della simulazione sono replicate, ogni volta modificando il coefficiente di discriminazione, che è fatto variare da 0,1 a 50[144], immaginando che le caratteristiche strutturali degli *item* possano essere fatte cambiare in modo da aumentarlo o diminuirlo. L'autrice considera, giustamente, anche altri fattori che concorrono ad aumentare la precisione di stima di un modello, legati alla distribuzione delle difficoltà degli *item*. Prevede che i valori delle difficoltà dei 16 *item* non siano sempre gli stessi, ma cambino la loro distribuzione rispetto a tre criteri: a) equidistanza dei valori o distanze differenziali, b) copertura di tut-

[143] A ciò ci sentiamo, però, di aggiungere che se un segnale di "disturbo" è introdotto in laboratorio, in modo controllato dallo sperimentatore, esso prende a far parte dell'insieme delle variabili operative, più che della componente di "errore".

[144] Ricordiamo che, sebbene nella versione *standard* del modello di Rasch il valore del coefficiente di discriminazione è convenzionalmente fissato pari a 1, esso può assumere valori diversi senza intaccare la natura parallela delle curve degli *item*: il requisito fondamentale è che il coefficiente sia di volta in volta *costante* per tutti gli *item* coinvolti nel test; se, all'interno dello stesso test, il coefficiente di discriminazione può *variare* da un *item* a un altro, allora passiamo al caso di un modello a due parametri.

to lo spettro dei valori dell'abilità o concentrazione in una specifica regione, c) presenza di una discontinuità (*gap*) tra due gruppi di valori tra loro concentrati o dispersi. Alla fine emergono sei diversi gruppi di *item*.

Zand Scholten passa quindi, come detto, a calcolare l'*errore di stima* del test delle abilità sulle risposte simulate a ogni insieme di *item* e, all'interno di ogni insieme, per ciascun livello di discriminazione, replicando ogni *step* 50 volte. La relazione diretta tra capacità di discriminare e sensibilità con cui un *item* riesce a fornire stime delle abilità contigue alla sua posizione, lascerebbe supporre che l'errore di stima decresca sistematicamente al crescere del coefficiente di discriminazione degli *item*, in ognuno dei sei scenari previsti dalla simulazione; al contrario, saremmo di fronte a una palese negazione del principio reclamato da Michell per fondare il paradosso di Guttman-Rasch, per cui "meno errore" significa sempre "misurazione più accurata".

Il risultato della simulazione è che vi sono dei casi (insiemi di valori di difficoltà degli *item*) in cui l'errore decresce fino a livelli del coefficiente di discriminazione compresi tra 2,5 e 5, per poi risalire più o meno vistosamente via via che quest'ultimo incrementa: in particolare ciò avviene nello scenario con valori degli *item* i) concentrati in una ristretta regione, ii) con una discontinuità e valori concentrati in due gruppi, iii) con valori non equidistanti e concentrati in una regione centrale. Ciò dimostrerebbe dunque, per Zand Scholten, che l'invarianza della relazione inversa postulata da Michell tra precisione ed errore può essere confutata. Sembreremmo di fronte a un vero e proprio contro-paradosso.

14.3.2 Accuratezza ed errore: richiamo al ruolo della funzione informativa

Innanzitutto notiamo, per inciso, che la prova di Zand Scholten, oltre che funzionare come grimaldello per disinnescare la validità del paradosso, rischia anche di risolversi in un involontario attacco al modello di Rasch. Dal punto di vista di Michell si potrebbe, infatti, ribattere che i) prima di sostenere che i risultati della simulazione sono una prova che a volte nella misurazione per aumentare la precisione bisogna aumentare anche l'errore, bisognerebbe aver dimostrato che l'attributo in questione è quantitativo e che si sta realmente misurando, mentre proprio questo è il nodo della questione che andrebbe sciolto e che ii) in questo modo si è oltretutto dimostrato che la sua stessa strut-

tura matematica permetterebbe alcuni rilevanti non-sensi (aumentano insieme precisione ed errore) che ne indeboliscono le pretese di misurazione.

Rimanendo, però, sui binari della riflessione di Zand Scholten (che ci pare sollevi aspetti assolutamente fondamentali), riflettendo sulle fattispecie della sua confutazione è possibile sostenere che non è necessario giungere alle medesime conclusioni nell'interpretazione del risultato. Esso può ritenersi tutt'altro che inaspettato per il modo stesso, peraltro esposto in modo trasparente e appropriato, con cui la simulazione è stata elaborata. Riprendiamo il concetto di funzione informativa introdotto nel capitolo sui modelli della IRT e avanziamo una possibile interpretazione dei risultati.

La funzione informativa è il reciproco della varianza della stima, per cui aumenta al decrescere della seconda. Essa ci consente di sintetizzare l'informazione sulla precisione della stima e sostituisce, in ambito continuo e probabilistico, il classico concetto di *affidabilità*. Nel caso dei modelli logistici a uno, due e tre parametri essa è massima in corrispondenza del livello di difficoltà dell'*item* e decresce rapidamente allontandosi da quel punto[145]. Per il modello a due parametri, poiché la varianza della stima è in ogni punto del *continuum* pari al reciproco del prodotto tra quadrato del coefficiente di discriminazione, probabilità di superare l'*item* e quella di non superarlo, allora la funzione informativa sarà:

$$I = a^2 * p * (1 - p) \qquad \text{(eq. 14.1)}$$

Nel caso del modello di Rasch in versione standard, dove il coefficiente di discriminazione costante è pari a 1, la funzione si riduce matematicamente al prodotto semplice tra le probabilità; nel caso in cui il valore si discosti da 1, pur rimanendo costante per ogni curva, si applicherà la formula 14.1 non semplificata. Se la discriminazione aumenta, la curva dell'*item* diventa più ripida nella sezione centrale ed è sensibile anche a sempre più piccole differenze di abilità (la probabilità aumenta o diminuisce molto più rapidamente allontanandosi dal punto di flesso della curva). Dal punto di vista della funzione informativa, ciò si traduce in un picco della curva sempre maggiore in termini

[145] Si rimanda alla sezione 5.4 per la giustificazione matematica del meccanismo che produce questa caratteristica, comunque intuitiva, della curva informativa.

assoluti di volta in volta che il coefficiente di discriminazione cresce, ma anche in una pendenza molto più accentuata.

La figura 14.1 mostra, banalmente, la distribuzione della funzione informativa lungo il *continuum* dell'abilità di cinque *item* ipotetici, di pari difficoltà ma discriminazione diversa. Come si può osservare, si tratta di cinque curve simmetriche e centrate, con altezze e pendenze diverse, che in basso s'intersecano tra loro. L'interpretazione è molto semplice: aumentando la discriminazione, cresce sempre di più la precisione (informazione) di un singolo *item* nello stimare soggetti con abilità vicina alla sua difficoltà, ma allo stesso tempo deteriora sempre più velocemente (pendenza) quella relativa ai soggetti più lontani, fino al punto in cui diventa meno informativa di *item* con coefficienti di discriminazione più bassi (intersezioni tra le curve).

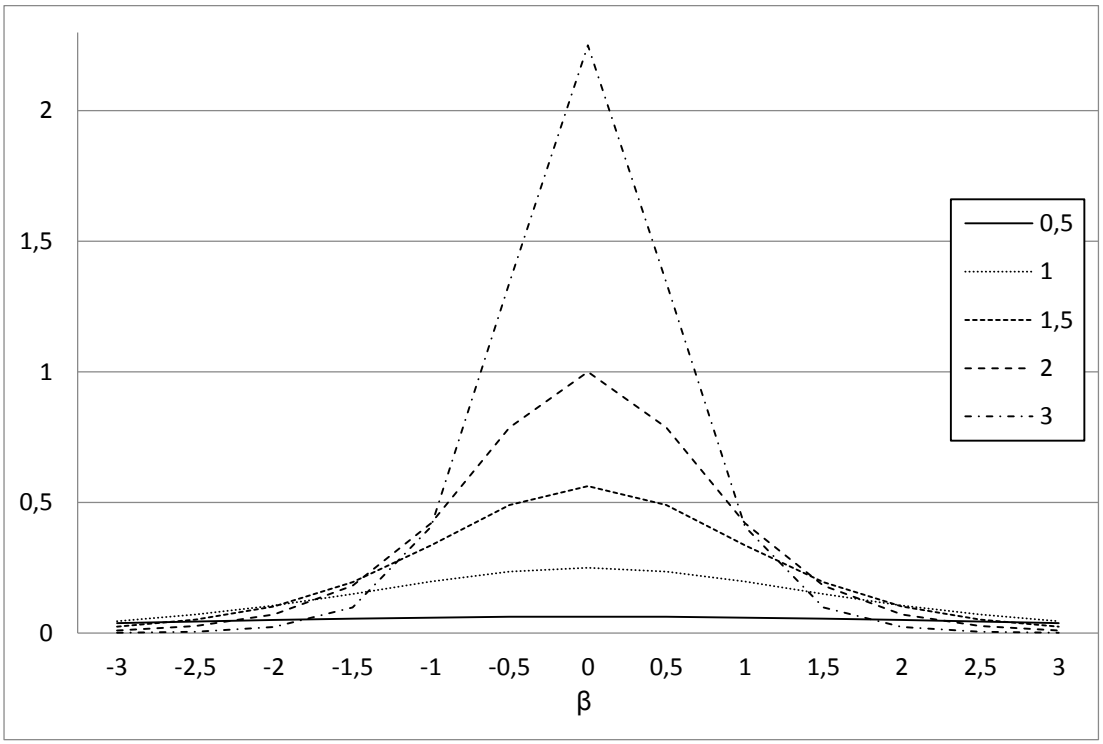


Fig. 14.1 – Esemplificazione di funzioni informative per *item* con difficoltà pari a 0 e coefficiente di discriminazione rispettivamente pari a 0.5, 1, 1.5, 2 e 3.

Insomma, ogni successivo incremento di precisione è sempre relativo a una regione progressivamente più ristretta di abilità, tanto che un *item* meno sensibile diventa comunque, alla distanza, più informativo rispetto a soggetti lontani da esso. Questo meccanismo, come accenna anche Zand Scholten, ha un effetto sull'errore di stima

quando si va a comporre il test[146]: una serie di *item* troppo sbilanciati rispetto alla distribuzione dei soggetti produrrà un errore più elevato, perché l'abilità dei soggetti collocati in regioni del *continuum* distanti dagli *item* sarà necessariamente stimata in modo approssimativo e poco preciso[147] e ciò tanto più siano discriminanti gli *item*.

14.3.3 Relazione diretta tra precisione ed errore dal punto di vista della funzione informativa

Per quale motivo, quindi, il risultato ricavato da Zand Scholten può anche essere interpretato in una direzione diversa da quella sostenuta dall'autrice? Il campione dei soggetti è stato estratto da una popolazione distribuita normalmente rispetto all'abilità, quindi in modo regolare e simmetrico lungo il *continuum*. In una situazione tale, una batteria di *item* concentrati in uno o più intervalli ristretti di valori, con la maggior parte del *continuum* dell'abilità lasciato scoperto, sarà destinata a produrre stime con maggiore errore standard (meno informative) per livelli di abilità distanti. Aumentando la sensibilità di questi *item* si è condotti al paradosso apparente: invece di migliorare, l'affidabilità del test alla lunga peggiora perché l'errore aumenta. Questo, lungi dall'essere una dimostrazione del fatto che, in via di principio, errore e precisione possono essere direttamente correlati, ci pare più la conseguenza di *item* distribuiti in modo disfunzionale rispetto alla natura del modello. Capiamo perché.

Da un punto di vista complessivo del test, quando gli *item* sono distribuiti disomogeneamente e/o concentrati, essi comportano a) un guadagno molto sostenuto in termini di riduzione dell'errore di stima per i soggetti (relativamente pochi) con abilità vicine alla difficoltà dell'*item* e b) una perdita per i soggetti relativamente più distanti. Aumentare la sensibilità (discriminazione) tende a massimizzare sia i benefici di a) che gli svantaggi di b). Se all'inizio il saldo è positivo e l'informazione del test cresce (cioè l'errore quadratico medio della stima diminuisce), innalzando ancora il livello di discriminazione la stima (sempre più precisa ma per regioni sempre più piccole e isolate) non sarà in grado di bilanciare l'effetto di b) su regioni sempre più grandi, che sarà anzi

[146] Ricordiamo che la curva d'informazione del test in un punto è la somma di tutte le curve d'informazione degli *item* in quello stesso punto.

[147] D'altra parte, se per qualche motivo si confida di conoscere già la distribuzione dell'abilità dei soggetti da rilevare, con il modello di Rasch si ha il vantaggio di poter massimizzare la qualità della stima utilizzando *item* la cui difficoltà si trovi nelle regioni corrispondenti.

amplificato a tal punto da invertire la tendenza, portando a un aumento dell'errore. Ecco giustificato il meccanismo che porta a questo risultato.

La tesi che possiamo avanzare, quindi, è che qui il problema sia legato alla costruzione dello strumento: come ampiamente ripetuto in letteratura, il modello di Rasch ha bisogno che i suoi *item* siano rappresentativi dell'intero *range* delle abilità, che coprano in modo omogeneo l'arco del *continuum* interessato dalla rilevazione. Non a caso, quando abbiamo affrontato il concetto d'informazione, abbiamo osservato che la situazione ideale sarebbe quella in cui la curva informativa del test sia tendenzialmente costante per tutti i valori, cioè composta di funzioni informative dell'*item* equidistanti e ravvicinate tra loro[148]. Una batteria di *item* le cui posizioni si discostino molto da quella ideale, è destinata a produrre "paradossi" come quello dell'errore e della sensibilità, perché intere aree del *continuum* sono prive di confronti precisi.

In questo caso, anche a noi sovviene una metafora presa a prestito dalle scienze fisiche per aiutare a visualizzare quanto sostenuto, più semplice di quella della risonanza stocastica. S'immagini un'asta di legno, incisa in alcuni tratti con una scala graduata al livello di sensibilità dei centimetri e in altri tratti senza alcuna graduazione (liscia). Evidentemente, se appoggiata a terra e accostata a una serie di oggetti, l'asta produrrebbe un errore di misurazione più elevato di un'asta che invece fosse graduata in tutta la sua lunghezza: rileverebbe l'altezza di alcuni oggetti con la sensibilità di un centimetro, ma per gli altri l'attribuzione di una misura avverrebbe necessariamente "a occhio". Ovviamente, tale errore di misurazione aumenterebbe al crescere degli spazi non graduati e/o all'aumentare degli oggetti da misurare la cui altezza fosse ricompresa in essi. Si pensi, poi, a cosa accadrebbe se progressivamente diminuissimo e restringessimo i tratti graduati, portando però la loro sensibilità al livello dei millimetri. E' ovvio che questi migliorino la precisione della misurazione, ma a costo di aumentare la platea di oggetti da "misurare a occhio", con un errore elevatissimo. Andando avanti, aumentando sempre di più la sensibilità (micrometri ecc.) ma restringendo i tratti graduati, fino quasi a rendere l'asta completamente liscia, non è difficile immaginare che alla fine quasi nessun oggetto sarebbe misurabile se non "a occhio" e che la precisione microscopica gua-

[148] Nel modello di Rasch, poiché il coefficiente di discriminazione è costante per tutti gli *item*, ogni curva d'informazione avrà un medesimo massimo. In particolare, se la discriminazione è pari a 1, l'informazione massima dell'*item* è pari a 0,25.

dagnata per intervalli sempre più piccoli non sarebbe in grado di compensare l'aumento dell'errore, risultando in una bassissima capacità informativa globale delle misurazioni.

Questo, con le dovute differenze (si tratta di una semplice metafora), è più o meno il meccanismo che coinvolge anche il nostro esempio con il modello di Rasch. Siamo abbastanza certi che, nel caso dell'asta graduata, nessuno sosterrebbe che da questo esempio si trae la conclusione che, in determinate circostanze, l'errore aiuti la precisione: piuttosto, si parlerebbe di un modo non appropriato di costruire uno strumento, che porta ad alcuni effetti bizzarri sotto precise circostanze. Lo stesso per il modello di Rasch: alla presenza di strumenti costruiti in modo disfunzionale e non appropriato alle sue caratteristiche, in certi casi molto particolari si osservano alcune conseguenze paradossali.

Per chiudere il cerchio del ragionamento, è interessante capire cosa accade nella simulazione di Zand Scholten quando ci troviamo nel caso-limite ideale in cui gli *item* sono equidistanti tra loro e coprono tutto il campo di variazione delle abilità del campione. La risposta è: in quel caso l'errore quadratico medio decresce sistematicamente al crescere della discriminazione degli *item*. In altre parole, se aumenta la sensibilità, allora diminuisce l'errore. Tornando per l'ultima volta alla nostra funzione informativa, significa che la curva del test è tendenzialmente costante e, all'avanzare della discriminazione, s'innalza progressivamente diminuendo l'errore.

Anche questo argomento di Zand Scholten ci pare quindi non risolutivo. Esso solleva questioni molto importanti e mette in luce alcune criticità del modello di Rasch che possono emergere quando ci si allontana dal quadro ideale che esso richiede. Allo stesso tempo abbiamo segnalato alcuni motivi per cui la simulazione proposta non riesce ad avanzare tutti gli elementi a nostro avviso necessari per mandare completamente a segno il colpo. L'autrice sottolinea che il suo obiettivo, in questo contesto, non è quello di sostenere che le scale prodotte dal modello di Rasch siano al livello della misurazione a intervalli, bensì di dimostrare come sia effettivamente possibile che un incremento della precisione possa convivere, anche in questo modello, con un aumento dell'errore. Si tratta di un tentativo di giustificare il parallelo con la risonanza stocastica, la quale ri-

marrebbe altrimenti solo un esempio interessante che però nessun legame ha con il nostro caso specifico.

Quello che si è tentato di mostrare è, in sintesi, che ciò non solo avviene in fattispecie molto specifiche e particolari, come segnala l'autrice, ma anche che non siamo in realtà di fronte a un aumento della precisione *tout court*, bensì a un risultato contingente di due tendenze in atto scomponibili: da un lato un miglioramento effettivo della precisione, di alta intensità ma su una porzione sempre più piccola della dimensione latente; dall'altro un peggioramento della precisione, d'intensità più bassa ma su una porzione sempre maggiore di tratto latente. Alla fine è il secondo effetto a prevalere, eccetto che non si faccia sì che gli *item* siano distribuiti in modo omogeneo lungo tutto l'arco dei valori. Tale questione ci pare doverla assegnare, piuttosto, al campo della corretta costruzione dello strumento di misurazione (da qui l'esempio dell'asta graduata) e non a una caratteristica intrinseca al rapporto stesso tra misurazione valida e strumento: è un punto di cui tenere conto, eccome, nella pratica dell'applicazione del modello e rientra nella lunga serie di aspetti che abbiamo sollevato e che aiutano a individuare e tenere distinti i risultati formali del modello dalle limitazioni del suo utilizzo pratico.

La precisione del modello di Rasch, come presunto modello di misurazione nell'insieme delle sue proprietà matematiche generali, non può prescindere dal fatto che ogni *item* apporta un contributo legato principalmente alla sua posizione e non all'intero *continuum*. Dato questo punto non aggirabile, quando gli *item* sono strutturati in modo tale da soddisfare tali requisiti, si osserva che l'errore diminuisce all'aumentare della discriminazione degli *item*. Quando quest'ultima raggiunge il suo massimo e sfocia nella condizione matematica teorica dell'infinito (cioè della precisione perfetta) ci troviamo sempre e comunque di nuovo con il modello di Guttman. Ed ecco che si ripresenta il paradosso di Rasch come illustrato da Michell.

A ciò possiamo aggiungere che, indipendentemente dal comportamento dell'errore di misurazione, questi esempi non risolvono la questione più generale posta da Michell: giustificare l'introduzione della probabilità per raggiungere una rappresentazione continua e quantitativa di attributi, senza aver dimostrato che essi siano davvero continui e quantitativi.

14.4 Precisione e probabilità: il parallelo con la fisica quantistica

Nel quadro del dibattito su probabilità, precisione ed errore ci pare rientri anche uno spunto affascinante quanto molto ardito, che vale la pena discutere, pur se maturato indipendentemente dal dibattito sul paradosso, soprattutto perché evocato anche da un autore come Andrich, uno dei più eminenti studiosi e sviluppatori dei modelli della IRT. Andrich oltrepassa i termini del discorso fin qui fatto, pur restando nell'ambito di un'analogia, ed evoca un parallelo tra lo sviluppo dei modelli probabilistici nelle scienze umane e quello di uno dei settori più avanzati delle scienze fisiche del XX secolo: la meccanica quantistica.

14.4.1 L'uso della probabilità

Andrich avanza una distinzione concettuale tra fattore aleatorio (come mezzo per riprodurre il funzionamento dei processi di risposta) e scarsa accuratezza dei dati (come errore di misurazione). Dopo, sottolinea che sarebbe proprio la grande accuratezza raggiunta nella misurazione dei tratti mentali a richiedere l'introduzione del primo nei modelli psicometrici. In particolare,

> la necessità di modelli probabilistici nelle scienze sociali non deriva da una mancanza di precisione dei dati, bensì il contrario; esso implica che gli scienziati sociali che lavorano al livello delle unità dei loro strumenti di misurazione hanno bisogno di modelli probabilistici solo perché stanno operando a livelli precisi rispetto alla variazione dei valori dei tratti che stanno misurando. Per supportare questa prospettiva, è stato osservato che, laddove gli strumenti di misurazione nelle scienze fisiche sono divenuti molto precisi, cioè al livello quantico, sono state invocate formulazioni probabilistiche. [Andrich 2003, 27].

Per cui, come nella fisica quantistica, anche nei modelli della IRT il fattore aleatorio sarebbe il frutto non di una qualche sfocatura nella capacità di descrivere un fenomeno, bensì di un'accresciuta accuratezza che aumenterebbe di molto la fedeltà del modello alla realtà, richiedendo una misurazione lungo la scala del continuo. L'analogia è, come detto, sorprendente e interessante e non è chiaro quali possano essere gli effettivi punti di contatto tra il settore più avanzato delle scienze in cui più sviluppata è la capacità di misurazione e una disciplina nella quale lo *status* della misurazione, come questo testo ha ribadito sin dall'inizio, è sottoposto a continue messe in discussione. Nel suo libro del 1955 "Natura e fisica moderna", Werner Heisenberg, che formalizzò il fondamentale principio d'indeterminazione che porta il suo nome, prova a riassumere i caratteri centrali della profonda svolta epistemologica della fisica quantistica, dandoci degli spunti molto interessanti per valutare la tesi di Andrich:

> La fisica newtoniana era costruita in modo che, conoscendo lo stato di un sistema a un certo tempo, si poteva calcolare in precedenza il movimento futuro del sistema stesso. (…) Se la parola "causalità" si interpreta in modo così stretto, si parla anche di "determinismo" e si intende dire che esistono leggi fisse di natura le quali, partendo dallo stato attuale di un sistema, determinano univocamente il suo stato futuro. [Heisenberg 1955/1957, cap. II]

Pensiamo al modello di Guttman. In esso il determinismo si manifesta in concetti come "scalabilità" e "riproducibilità": conoscendo i punteggi di una serie di soggetti e quelli di una serie di *item*, si può stabilire con assoluta precisione *quali* soggetti hanno risposto positivamente (o correttamente) a *quali item*. Nel modello probabilistico di Rasch, invece, il punteggio permette di stimare la posizione dei soggetti e degli *item* su un'ipotetica scala metrica, in quanto statistica sufficiente, ma non rende possibile stabilire quali soggetti hanno superato quali *item*[149]. Il determinismo nei modelli psicometri-

[149] Michell [2000] nella sua polemica contro l'approccio del "pitagorismo", tocca questo tema quando afferma che lo schema di risposta è più *fondamentale* del punteggio: i) perché dal primo si ricava il secondo e non viceversa; ii) perché il processo psicologico causa lo schema di risposta, non il punteggio. In generale, lo schema di risposte è un ordine parziale, non necessariamente una struttura quantitativa, a differenza invece dei punteggi del test.

ci, quindi, è tale principalmente rispetto alla "struttura" rilevata nei dati e la causalità riguarda la coerenza delle risposte riguardo a un supposto attributo latente che le avrebbe determinate in quel momento, sempre al netto di altri fattori d'influenza. Si rifletta: è sempre possibile immaginare che un soggetto supererebbe un qualsiasi ulteriore *item* più facile di quelli che ha già superato o che non supererebbe uno qualsiasi più difficile di quelli che non ha superato. In qualche modo, quindi, si può assumere di poter prevedere lo stato di un soggetto su un *item* successivo, così come il determinismo meccanicistico della fisica newtoniana permette di stabilire lo stato di un sistema in un momento futuro[150]. Ci sono però delle difficoltà: per esempio, l'ordine di soggetti e *item* nel modello deterministico di Guttman è accertato congiuntamente e globalmente, test per test e non sappiamo se inserendo o no altri *item* avremmo avuto gli stessi esiti, né potremmo confidare in risultati confrontabili per campioni di soggetti diversi, né sapremmo se un soggetto avrebbe superato un nuovo *item* la cui difficoltà si collocasse tra quella del più "difficile" accettato e quella del più "facile" rifiutato ecc. Insomma, nella pratica possiamo osservare una certa differenza tra il determinismo del modello psicometrico che riguarda principalmente la sua struttura logica interna e quello meccanicistico e nomotetico di stampo newtoniano.

Prosegue Heisenberg:

> Non si è tuttavia rinunciato teoricamente al determinismo fino alla celebre scoperta di Max Planck (…) che un atomo radiante non perde la sua energia in modo continuo, ma discontinuamente, a scatti. Questa perdita di energia discontinua e a scatti porta (…) all'ipotesi che l'emissione di radiazioni sia un fenomeno statistico. (…) *Effettivamente la teoria dei quanti costringe addirittura a formulare le leggi proprio come leggi statistiche* e ad abbandonare, anche teoricamente, il determinismo. (…) La divergenza dalla fisica precedente si rileva nelle cosiddette "relazioni di indeterminazione". Si è constatato che non è possibile indicare simultaneamente, con un grado qualunque di esattezza, la posizione e la velocità di una particella elementare. [ibidem]

[150] Ciò richiederebbe che atteggiamenti, valori, opinioni di esseri coscienti e non fungibili siano sottoposti a leggi immutabili e non influenzati da nient'altro nel tempo che dall'attributo da misurare. Quanto questo quadro sia verosimile si lascia valutare al lettore.

Qui è spiegato il motivo del passaggio da uno schema di spiegazione deterministico a uno statistico. La grande rivoluzione teorica della fisica nasce dall'esigenza di ricostruire un quadro concettuale che riesca a spiegare il comportamento delle particelle, altrimenti non interpretabile secondo la logica della meccanica newtoniana. Semplificando in modo estremo, l'energia si sposta in "pacchetti" *discreti*, la descrizione dei cui "salti" da un livello a un altro richiede un'illustrazione probabilistica. Curiosamente si può notare che, al contrario, l'obiettivo dell'introduzione della probabilità nei modelli IRT è di passare da punteggi discreti a stime metriche lungo dimensioni assunte come *continue*. Appare qui chiaro quanto il parallelo tra l'uso del fattore aleatorio in campo fisico e psico-sociale esponga a non poche difficoltà: molto diverse sono le esigenze e le basi su cui esso poggia e la logica di studio del rispettivo oggetto d'interesse. Appare un po' forzato, o come minimo prematuro, sostenere che i modelli psicometrici stiano percorrendo la stessa strada della fisica quantistica.

14.4.2 Indeterminazione e complementarità

Heisenberg introduce un altro aspetto interessante per la nostra riflessione:

> Un'altra formulazione è stata coniata da Niels Bohr, che ha introdotto il *concetto della complementarità*. Egli intende con questo che diverse immagini intuitive, con cui noi descriviamo sistemi atomici, sono sì adatte per certi esperimenti, ma si escludono reciprocamente. (…) Questi diversi modelli sono quindi giusti quando li si utilizzi al posto giusto, ma si contraddicono fra loro e si chiamano, perciò, reciprocamente complementari. L'indeterminazione da cui ognuna di queste immagini è affetta e che viene espressa mediante la relazione di indeterminazione, basta appunto ad evitare contraddizioni logiche fra le diverse immagini. (…) *La conoscenza incompleta di un sistema deve essere una componente essenziale di ogni formulazione della teoria quantistica*. Le leggi quantistiche devono essere di tipo statistico. [ibidem]

Se è legittimo nutrire dubbi sull'analogia tra fisica e scienze sociali rispetto ai modelli probabilistici, è altrettanto vero che dobbiamo fare tesoro di concetti come quello di complementarità e indeterminazione, i quali costituiscono un monito rispetto alla rappresentazione che della ricerca in ambito fisico è spesso fatta dagli studiosi sociali che ne vogliono fare un modello universalmente valido di scienza. Quello che sostiene Heisenberg è che non esiste un solo modo corretto di rappresentare un sistema fisico a livello di particelle: modelli tra loro in contraddizione possono dimostrarsi, in situazioni diverse, modi altrettanto efficaci di rappresentare certi sistemi in certe condizioni. Ma c'è di più, perché questo chiama in discussione eventuali interpretazioni ingenue del realismo empirico, laddove si dovesse pensare che per le scienze fisiche "realtà" significhi un oggetto di ricerca concepito come qualcosa "là fuori", indipendente dalle operazioni di misurazione, determinato da rapporti precisi che è compito dello scienziato *scoprire* in modo neutrale e testare in modo rigidamente assiomatico. Heisenberg e Bohr costatano che l'apparato strumentale (cioè l'intervento dello scienziato) interagisce con il sistema atomico e che il punto di vista da cui si osserva il fenomeno (le condizioni sperimentali determinate dal ricercatore) può dar vita a "immagini" diverse ma altrettanto legittime, tutte vincolate da una qualche forma d'indeterminatezza. Paradigmatico nella storia della fisica è il dibattito sulla natura delle particelle:

> I diversi esperimenti che dimostrano sia la natura ondulatoria, sia quella corpuscolare della materia atomica, ci costringono, con i loro paradossi, a formulare delle leggi statistiche. Nei processi in campo macroscopico questo elemento statistico della fisica atomica non ha in generale importanza, perché nel processo macroscopico deriva dalle leggi statistiche una probabilità così elevata, da permetterci di dire che il processo è, praticamente, determinato. [ibidem]

Quest'ultimo passaggio ci ricorda che oltre alla coesistenza di varie "immagini" per lo stesso sistema, possono coesistere anche diverse "fisiche" a scale diverse di grandezza: nella vita di tutti i giorni (a livello macroscopico e a velocità non tendenti a quella della luce) la fisica deterministica newtoniana *funziona* correttamente, è *utile*, riesce a

svolgere il suo compito perché i fenomeni microscopici probabilistici si compongono dandoci la certezza sostanziale di cosa accadrà. A livello di particelle o a velocità vicine a quella della luce, invece, meccanica quantistica e teoria della relatività diventano strumenti fondamentali[151].

Questa analogia offre argomenti potenzialmente a favore di posizioni diverse. Infatti, da un lato potremmo chiederci: se la convivenza di "immagini intuitive" tra loro complementari vale per la fisica, perché non dovrebbe valere a maggior ragione per le scienze sociali[152]? Ciò forse aiuterebbe a sdrammatizzare il fatto che modelli diversi producano "immagini" diverse. Inoltre, questo rafforzerebbe il punto di vista di chi vede nei modelli probabilistici un modo legittimo per organizzare organicamente una serie d'ipotesi e assunti, controllandone la validità attraverso il confronto con il materiale empirico, senza necessariamente dover testare prima assiomi deterministici troppo rigidi che stabiliscano una volta per tutte la natura degli attributi. Senza contare che sarebbe legittimato e riconosciuto come elemento fondamentale l'intervento del ricercatore, abbandonando la chimera di una presunta oggettività *super partes*: un modello, costruito come strumento matematico da uno scienziato, tiene insieme e fa emergere anche questo aspetto d'inevitabile soggettività.

Dall'altro canto, il fatto che esista una sorta di "principio di adeguatezza" di un modello rispetto a situazioni diverse, è proprio quello che spinge studiosi come Michell a invitare all'uso di scale ordinali per rappresentare le osservazioni: per quale motivo voler imporre scale metriche su dati empirici dallo *status* così incerto? Qual è l'utilità di ricercare un livello di precisione e sensibilità così sproporzionato rispetto all'ambizione delle teorie in uso? Soprattutto, le varie applicazioni empiriche di certi modelli hanno corroborato veramente le loro ipotesi in ambito umano oppure la generale mancanza di convergenza tra modelli e dati è stata indulgentemente "coperta" da una certa tendenza alla "negazione della falsificazione"? [Heene 2013]

[151] Un ingegnere che progetta un treno non ha a che fare con la meccanica quantistica molto di più di un appassionato di meccanismi di orologi, né ha bisogno della teoria della relatività perché, per quanto rapida possa essere la motrice che dovrà ideare, la sua velocità sarà sempre più vicina a quella di una tartaruga che a quella della luce.

[152] Per un esempio di elaborazione teorica che ricerca un parallelo tra le immagini della società e quelle della fisica quantistica, si rimanda a Zohar e Marshall [1994].

L’analogia dei modelli IRT con la fisica quantistica ci pare, tirando le somme, fermarsi molto semplicemente a un qualche utilizzo della probabilità. Non riguarda meccanismi comuni o percorsi assimilabili. Stiamo parlando di situazioni molto diverse, in contesti di ricerca assai distanti e a livelli di sviluppo incomparabili. Appare complicato sostenere che, poiché alcuni modelli psicometrici utilizzano un fattore aleatorio e siccome anche la fisica quantistica utilizza modelli probabilistici, allora i primi sono paragonabili alla seconda e ne posseggono alcuni caratteri particolari, nello specifico la grande precisione raggiunta nella misurazione. Oltretutto andrebbe aggiunto che nel campo della fisica l’adozione della statistica non è il portato (solo e soprattutto) di una maggiore precisione nella misurazione, ma è il frutto di un vero e proprio sconvolgimento nella concezione stessa della causalità e dall’introduzione del principio d’indeterminazione [Barrett 2008], che viene da molto lontano e che, giova ricordarlo, avviene comunque nell’ambito di una scienza rigorosamente sperimentale. Concediamo quindi che tale accostamento sia, al massimo, un buon auspicio per il futuro delle scienze psico-sociali, ma non può fondare una confutazione giustificata degli argomenti paradossali del passaggio dal determinismo alla probabilità.

CONCLUSIONI

Misurazione: un concetto complesso

La riflessione compiuta in questo testo ha necessariamente preso le mosse dal concetto di *misurazione*: solo tracciandone i contorni è possibile riflettere sulla capacità di un modello di produrre un tipo d'informazione che possa essere riconosciuta come una *misura*. Il primo aspetto emerso con grande evidenza è che si tratta di un concetto problematico, complesso e su cui non vige una chiara concordanza intersoggettiva nelle comunità degli scienziati umani.

Esiste una concezione "classica" della misurazione come assegnazione di valori a proprietà continue quantitative in conformità a rapporti tra grandezze, con questi ultimi delineati secondo la tradizionale definizione euclidea, molto adatta alle caratteristiche delle procedure, degli attributi e delle teorie nell'ambito della fisica. Tale visione, incardinata poi nell'interpretazione rappresentazionale sviluppata tra la fine del XIX e l'inizio del XX secolo, ha costituito nella sua versione realista lo standard teorico del concetto di misurazione per lungo tempo. Ed è proprio in conformità a questa impostazione che le scienze psicologiche e sociali, a metà del XX secolo, sono state accusate di utilizzare abusivamente l'etichetta di "misurazione" per le procedure di rilevazione e stima in psicofisica, non avendone dimostrata la legittimità attraverso l'accertamento della natura quantitativa dei costrutti in oggetto. E' in questa situazione che si è fatto spazio un approccio revisionato, quello di Stevens, con la sua teoria delle scale, ispirato da una concezione della misurazione operazionalista che ha spostato il baricentro dalla struttura empirica delle proprietà alle operazioni utilizzate per la costruzione delle scale e alle manipolazioni ammissibili su di esse.

Per alcuni questo è stato "l'uovo di Colombo" che ha permesso piena legittimazione alla pratica di costruzione di variabili-scala nelle scienze umane come forma di misurazione, anche con riferimento allo studio di atteggiamenti e dimensioni mentali latenti in genere; per altri, tuttavia, ciò ha significato un indebito allargamento del concetto di misurazione, che ha reso tutto misurabile e lasciato evaporare i criteri di controllo della rappresentabilità di una struttura empirica su una formale a intervalli (e quindi i criteri per l'accertamento del livello di misurazione).

Abbiamo, quindi, riconosciuto che la determinazione di cosa *significhi* misurare e cosa *sia* misurazione è una questione tutt'altro che lineare e priva di controversie, né tantomeno risolta, bensì dipendente dal tipo di definizione teorica che se ne voglia dare

e dai caratteri dell'oggetto di ricerca. In particolare, si è osservato come almeno tre degli aspetti coinvolti nelle procedure di misurazione – e cioè la struttura della proprietà empirica, le procedure di costruzione delle scale formali, le statistiche e le trasformazioni invarianti ammissibili al livello di misurazione raggiunto – si combinano per stabilire, secondo i vari punti di vista, se gli attributi teorici latenti delle scienze psico-sociali siano misurabili o meno in via di principio, se il concetto di misurazione utilizzato debba essere derivato da quello classico realista o da quello rielaborato in chiave rappresentazionale-operazionalista, e se le regole di assegnazione siano da interpretare prevalentemente come funzioni generate dalle strutture (in cui la realtà determina vincoli precisi alla rappresentazione formale) ovvero funzioni che generano e definiscono le strutture (in cui sono le procedure di costruzione delle scale a costituirne i caratteri formali).

Il successivo richiamo alle pratiche della costruzione di scale (*scaling*) negli studi sugli atteggiamenti ha fornito qualche altro cenno teorico e tecnico su come il tema è stato inquadrato e portato ad applicazione da alcuni studiosi nel tempo. Tra gli esempi classici discussi, il modello di Guttman ha avuto importanza, non solo per il seguito della trattazione (essendo considerato un precursore deterministico del modello di Rasch) ma anche per la sua caratterizzazione rigorosamente ordinale, che interpreta il concetto di *scalabilità* in modo del tutto indipendente dalla ricerca d'intervalli. La riflessione di Thurstone è stata altrettanto utile e fondamentale per chi si stava accingendo a riflettere sulla possibilità di costruire scale quantitative di atteggiamenti, giacché si tratta di uno dei primi e dei più celebri tentativi, nella tradizione psico-sociale, di definire e dare un fondamento robusto al concetto di "intervalli uguali" nelle scale di test mentali. Infine, il modello *unfolding* di Coombs ha rivestito un profilo di largo interesse poiché, pur nella sua relativa semplicità e parsimonia di assunti, permette di indagare le condizioni di possibilità che rendono immaginabile utilizzare dati squisitamente ordinali (quali quelli generalmente raccolti nelle scienze umane) per ricavare informazioni anche sull'ordine delle differenze tra posizioni lungo un *continuum*, costituendo un tentativo di porre un ponte tra scale ordinali e scale a intervalli.

Il modello di Rasch: punti di forza e interrogativi

La Teoria del Tratto Latente (TTL) e i modelli probabilistici della *Item Response Theory* (IRT) si presentano come un'alternativa ai modelli della Teoria Classica dei Test e, in generale, deterministici. Coerentemente con la constatazione che il concetto di misurazione ha specificazioni diverse in base alle prospettive teoriche adottate e agli assunti selezionati per giustificarne le procedure, anche il *focus* che abbiamo scelto per la discussione dei caratteri e delle proprietà dei modelli dicotomici della IRT si è attenuto a tale principio. Per questo, più che concentrarci sulle procedure tecniche di stima dei parametri, il punto di vista adottato ha avuto a che fare in modo più diretto con la discussione preliminare dei rapporti epistemologici fondativi che è necessario ipotizzare (tra gli attributi, tra gli attributi e gli indicatori operativizzabili e tra entrambi e i modelli formali) per giustificare la rappresentazione avanzata dai modelli. In tal senso, sono stati esposti gli assunti teorici fondamentali sul rapporto tra base empirica e modello e analizzate le caratteristiche delle curve che descrivono il rapporto tra dimensione latente e probabilità di superare un *item*, evidenziando la connessione tra il significato dei parametri e i caratteri empirici degli *item*. Prima di passare al modello di Rasch a un parametro, l'illustrazione dei modelli dicotomici a tre e due parametri è stata l'occasione di riflettere su come ognuno di essi, possedendo una certa struttura formale matematica dotata di particolari proprietà, offra necessariamente una diversa interpretazione del rapporto tra *item* e meccanismi di risposta del soggetto, confermando la natura profondamente densa di teoria di ognuno di essi.

Il modello di Rasch a un parametro, nella sua semplicità ed eleganza, teorizza un rapporto funzionale continuo tra la proprietà latente e la probabilità di risposta a un *item* che di tale proprietà è indicatore; tale funzione è basata esclusivamente sulla differenza tra la proprietà del soggetto e quella dell'*item*, posti su uno stesso *continuum* perché medesima è la dimensione latente, come nel caso del modello di Guttman e di quello di Coombs. In più, il modello di Rasch utilizza la natura continua delle curve di probabilità per costruire stime dei parametri di individui e *item* lungo scale a intervalli, invarianti rispetto a trasformazioni lineari: questa caratteristica costituisce il più importante elemento di dibattito e riflessione. Allo stesso tempo, il modello è molto rigoroso nella sua struttura formale poiché, ipotizzando che le curve degli *item* descrivano tracciati paralle-

li, permette un'interpretazione cumulativa delle risposte e il rispetto del principio di scalabilità, già proprio del modello di Guttman, di cui può essere interpretato come una forma probabilistica e metrica. Di queste caratteristiche abbiamo detto che costituiscono una base formale e matematica necessaria per paragonare i risultati di un modello di *scaling* alle pretese quantitative della misurazione in senso stretto.

Le proprietà fondamentali del modello di Rasch, da non confondersi con gli assunti preliminari, sono assolutamente strategiche in questo senso e sono in buona parte esclusive rispetto agli altri modelli IRT dicotomici con più parametri, che più facilmente si adattano ai dati ma che perdono alcune peculiari capacità del modello più semplice. Un punto fondamentale nella nostra illustrazione critica, talvolta non abbastanza rimarcato in letteratura, è stato la distinzione delle proprietà matematiche dei parametri del modello formale da quelle della base empirica della stima, mettendo in guardia da un'impropria e illegittima identificazione automatica delle prime con le seconde: solo nel caso di un adattamento molto soddisfacente del modello ai dati è possibile pensare di estendere, con una qualche appropriatezza, i caratteri delle stime dei parametri dalla sfera formale del modello a quella dei dati che le hanno prodotte. Resta quindi aperta la questione del passaggio dall'attributo empirico alla rappresentazione formale.

Tra le proprietà principali, ci sono quelle riguardanti la dimensionalità. Il modello di Rasch consente di rappresentare i parametri di soggetti e *item* su uno stesso *continuum* comune, attraverso una metrica lineare ottenuta da una trasformazione logaritmica e, soprattutto, utilizzando una "unità di misura" (il *logit*, differenza tra il valore dell'abilità del soggetto e quella dell'*item*) che rende in via di principio confrontabili le stime. Riflettendo su questo aspetto, ci siamo imbattuti in alcune importanti differenze, procedurali e sostanziali, che distinguono questa fattispecie di "unità" dal concetto classico di "unità di misura" di proprietà estensive ma non solo. Nel caso del *logit* si è rilevato il ruolo che esso gioca nella calibrazione e nel confronto tra scale, ma anche che esso non possiede il carattere di standard ripetibile indipendente dalla singola misurazione, né una capacità esplicativa esterna rispetto al significato della scala, essendo frutto di un rapporto strutturale interno a essa. Se parliamo di un'asta lunga tre metri (grandezza estensiva) o di un recipiente di acqua a venti gradi centigradi (grandezza intensiva), abbiamo dei riferimenti precisi che rendono esplicito il significato della misura: nel primo caso è quasi auto-evidente per la natura combinatoria del procedimento, in cui

l'accertamento stesso della grandezza (la combinazione di tre unità di misura) costituisce di per sé un modo per decrittarne il significato; nel secondo caso, perché sono noti e universalmente riconosciuti alcuni punti notevoli che connotano la scala (zero gradi, cento gradi ecc.). Parlare invece di due *logit* come "due volte il logaritmo del rapporto tra la probabilità di superare un *item* e quella di non superarlo" o come differenza di due unità tra abilità e difficoltà, non fornisce gli stessi elementi di esplicitazione. Il *logit* può dunque essere interpretato come un'unità di scala utilizzata nella segmentazione del *continuum*, più che un'unità di misura in senso stretto, utilizzata nelle procedure (appunto) di misurazione. Da questo punto di vista è più simile ai gradi della temperatura su una scala a intervalli Celsius o Kelvin, ma con la differenza che non esistono per il momento dei punti notevoli con un significato intersoggettivamente accettato, in grado di fissare la scala; senza contare che della temperatura, per continuare l'esempio, possono essere date anche altre interpretazioni fisiche, come il livello di movimento molecolare ecc.

Sulle problematiche della metrica, della calibrazione e dell'ancoramento dei test, connesse a quelle del significato e della validità della misurazione, ci siamo soffermati con più attenzione. In particolare, la natura non estensiva e manipolabile delle proprietà rappresentate dal modello di Rasch pone con più evidenza il problema del ruolo della teoria nell'interpretazione dei risultati: se, come alcuni autori hanno ritenuto, si dovrebbe accettare la natura fondamentale e additiva del tipo di misurazione in questione, il tema della calibrazione e dell'ancoramento del test svolge un ruolo più tecnico e arbitrario, rispetto al controllo dell'invarianza e dell'adattamento del modello ai dati, perché il significato è assegnato dalle procedure stesse di rilevazione e di allocazione dello "zero" sulla scala; diversamente, se il giudizio di uno scarso sviluppo delle teorie sugli attributi conduce a ipotizzare un livello di misurazione più incerto, allora il problema dell'invarianza e della calibrazione assume una valenza diversa ed emerge il tema dell'interpretabilità e del significato dei punteggi e del loro legame con fattori esterni al processo di rilevazione. Quest'ultima posizione poggia tra l'altro sull'osservazione che fattori di contesto e di significato degli *item* influenzano i processi di risposta a questi ultimi e rendono più complesso ammettere la loro reciproca indipendenza, postulata dal modello come assunto fondamentale. Insomma, torna anche qui la questione problema-

tica del rapporto tra le procedure di rilevazione, i modelli formali e la definizione teorica degli attributi empirici.

La vera proprietà decisiva del modello è quella delle statistiche sufficienti: il numero di *item* superati da un soggetto contiene tutta l'informazione relativa alla sua posizione sul *continuum* e il livello di proprietà di un *item* è indipendente dalla distribuzione della proprietà stessa tra i soggetti; viceversa, vale anche che il numero di soggetti che superano un *item* fornisce già tutta l'informazione sulla sua posizione sul *continuum* e il livello di proprietà di un soggetto è indipendente dalla distribuzione della proprietà stessa tra gli *item*. Tra le molte altre (indipendenza dal campione di soggetti e dal *pool* di *item* ecc.), la sufficienza delle statistiche ha come conseguenza anche l'invarianza dei parametri, che rappresenta la più importante caratteristica del modello.

L'invarianza dei parametri degli *item* dai soggetti e dei soggetti dagli *item*, introduce il concetto cruciale di *oggettività specifica*, la caratteristica formale più dirompente del modello di Rasch. L'invarianza, infatti, è considerata il presupposto indispensabile di qualsiasi misurazione in senso stretto, perché in linea teorica permette di ottenere gli stessi risultati indipendentemente dallo strumento o dal campione usati contingentemente; a sua volta, è la condizione per sperare di poter dare ai propri risultati un rilievo di generalizzabilità. Se voglio sperare di poter misurare un certo attributo e non altro da quello, è necessario che esperimenti o prove o osservazioni diverse convergano in un medesimo risultato, altrimenti si può sospettare che ciò che stiamo rilevando sia ogni volta qualcosa di diverso o un differente aspetto dello stesso fenomeno. L'invarianza nel modello di Rasch, per soggetti e *item*, può essere dimostrata attraverso varie strade, alcune delle quali abbiamo discusso, da punti di vista diversi (algebrico, assiomatico, grafico ecc.). Come per gli altri aspetti, anche l'invarianza e l'oggettività specifica sono risultati definitivi solo per il modello formale, poiché se la base empirica non rispetta gli assunti di unidimensionalità e indipendenza e se il livello di adattamento non è soddisfacente, allora le stime potrebbero essere solo un'immagine fittizia, inficiando qualsiasi illusione di oggettività.

Tre sono gli ordini di riflessione emersi dalla discussione dei caratteri del modello di Rasch: i) le proprietà del modello appartengono alla sua sfera formale e, allo stato dell'arte delle procedure di rilevazione e delle teorie sugli attributi, non possono essere

attribuite automaticamente e con certezza alle stime sulla base empirica: ciò che il modello di Rasch ha la forza di rappresentare, non necessariamente è ciò che viene, in effetti, rappresentato; ii) esistono approcci epistemologici distinti che interpretano in modo diverso la relazione tra assunti, proprietà del modello e base empirica, lasciando assegnata al dibattito teorico la questione controversa del livello di misurazione in atto; iii) nella sfera empirica, il rispetto degli assunti del modello (e per quali attributi) è un tema aperto che chiama in causa i caratteri costitutivi dell'oggetto di ricerca delle scienze psicologiche e sociali, mettendo in discussione la possibilità che un modello formale possa interpretarli.

Fondare la misurazione: Rasch e la misurazione additiva congiunta

Partendo da queste riflessioni, abbiamo dedicato spazio al dibattito sul fondamento del modello di Rasch come fattispecie di misurazione, sulla pretesa di produrre scale a intervalli e sulla struttura empirica degli attributi che intende misurare. Se le scienze psicologiche e sociali, come alcuni autori sostengono, possono ambire a raggiungere un livello di misurazione a intervalli analogo a quello delle scienze naturali e se l'obiettivo è di rivendicare risultati come quello dell'invarianza delle stime e dell'oggettività specifica, generalizzandoli oltre una dimensione specifica e contingente, allora s'impone una riflessione sulla natura degli attributi, sul suo accertamento e sui vincoli che impone ai processi di misurazione.

La Teoria della Misurazione Additiva Congiunta (ACM) individua e sviluppa una serie di condizioni il cui rispetto è necessario per dimostrare che due attributi siano quantità continue additive, a partire dal controllo dei rapporti d'ordine esistenti tra i valori di un terzo attributo costituito dalla combinazione non interattiva dei primi due. Tale teoria mostra, in modo rigoroso e con approccio assiomatico, che è possibile individuare la natura additiva di certe strutture empiriche, anche quando su esse non è possibile un procedimento di concatenazione, come le proprietà di tipo intensivo e, in particolare, quelle psicologiche.

Confrontando in modo sistematico gli assiomi ACM con la matrice delle stime e delle probabilità del modello di Rasch, si è mostrato perché, da un punto di vista mate-

matico, il secondo sia senza dubbio una fattispecie di struttura additiva congiunta, incorporando i primi. Questa caratteristica permette di interpretare il modello come fattispecie probabilistica di misurazione congiunta, facendo sì che possa potenzialmente rintracciare strutture quantitative, misurabili in modo sensato su scale a intervalli. Ciò farebbe del modello di Rasch uno degli strumenti più semplici ma allo stesso tempo completi ed efficaci per puntare a questo scopo nelle scienze umane.

Tale risultato fa i conti con alcuni aspetti che ne mettono in dubbio la portata. Il modello di Rasch è basato su una logica probabilistica, mentre gli assiomi dell'ACM su una deterministica. Gli assiomi testati sul modello di Rasch riguardano le stime prodotte dalla particolare curva di probabilità adottata, richiedendo una successiva analisi del livello di adattamento; ma l'analisi del *fit* pone alcuni seri interrogativi, non potendo essere sempre garanzia che i dati provengano effettivamente da strutture quantitative. Infine, vi sono tutti gli aspetti ricordati e legati al rapporto tra le strutture empiriche (soprattutto come definite in seguito agli sviluppi della moderna Teoria Rappresentazionale della Misurazione (RMT), nel cui alveo l'ACM è stata sviluppata) e il modello formale. I caratteri teorici degli attributi mentali e il modo in cui un test è costruito, con i criteri per la selezione di un certo *pool* di *item*, costituiscono ancora per molti autori aspetti di debolezza e differenza rispetto, invece, al controllo dell'additività su grandezze estensive e manipolabili.

Nel dibattito recente, due posizioni in particolare tendono a negare in modo più netto il legame tra ACM e modello di Rasch. La prima, esemplificata da Kyngdon, parte da un'ottica rigorosamente legata ai principi della RMT e si fonda principalmente sulla critica al tipo di verifica degli assiomi nel modello di Rasch, nel quale sulla struttura formale dei numeri reali sarebbe rappresentata una struttura altrettanto formale e numerica: le stime dei parametri e le probabilità. In questo modo non avremmo nessuna struttura empirica sulla quale testare gli assiomi ACM per dimostrare la natura additiva e quantitativa degli attributi, con ciò impedendo di accertare se effettivamente la dimensione latente si possa o no considerare una struttura additiva rappresentabile su una scala numerica a intervalli. Pur condividendo parte di queste riserve, si è visto che altri argomenti possono ridimensionare il ragionamento di Kyngdon, soprattutto rispetto alla sua concezione molto rigida e ristretta di ciò che sarebbe un sistema empirico direttamente osservabile. E' il caso di Sherry e della sua definizione di "sistemi quasi-

empirici" utilizzati nei processi di rappresentazione, che ci sembra rendere conto in modo più accorto della complessità del concetto di "relazioni empiriche" e dell'interazione tra teoria, formalizzazione e base empirica, allentando la forza dell'argomento di Kyngdon.

Vi è poi la posizione rappresentata da Michell, realista e vicino a una concezione tradizionale della misurazione. Egli punta dritto al cuore del nodo epistemologico: ritiene l'ACM uno dei possibili strumenti utili a porre finalmente termine all'incertezza che circonderebbe le teorie sugli attributi psicologici, che nella pratica dei modelli psicometrici sarebbero comunemente "misurati" senza però che sia quasi mai accertata la loro natura quantitativa e continua, data per scontata sulla base di una conoscenza convenzionale intrisa di pre-giudizi che egli definisce "pitagorici". Per lui, se è pur vero che il modello di Rasch mostra le caratteristiche strutturali degli assiomi ACM per distinguere strutture di tipo quantitativo da quelle semplicemente ordinali (soprattutto la cosiddetta gerarchia delle cancellazioni), esso si discosta comunque dalla logica dell'ACM: giunge al risultato finale non attraverso il controllo delle relazioni d'ordine dell'attributo congiunto, bensì tramite l'utilizzo di una specifica funzione che assume l'intervento di un ipotetico fattore aleatorio, giustificando in tal modo la relazione tra gli attributi e la probabilità di risposta. Per far ciò, si dà per scontata tutta una serie di aspetti (compresa la stessa natura quantitativa e continua della proprietà latente) che andrebbe invece testata empiricamente.

L'aspetto più particolare della sua critica si concentra proprio sul fatto che l'introduzione arbitraria dell'errore casuale (sotto forma di fattore aleatorio e di probabilità) sarebbe sufficiente a giustificare la pretesa dei modelli probabilistici di produrre misurazioni su scale a intervalli. Al contrario, l'esito di tale inserimento può addirittura far scaturire esiti logicamente sorprendenti, come egli mostra formalizzando quello che definisce come "paradosso dei modelli di Guttman e Rasch": se aumentiamo il livello di discriminazione nel modello a un parametro, miglioriamo la capacità dell'*item* di distinguere tra soggetti sempre più vicini, accresciamo cioè la precisione del confronto; se portiamo all'infinito tale livello, siamo ricondotti alla fattispecie del modello deterministico di Guttman, che produce scale di livello ordinale. I termini del paradosso sono evidenti: se ammettiamo che un aumento della precisione dovrebbe sempre migliorare la nostra capacità di misurare, com'è possibile che nel passaggio dal modello di Gutt-

man a quello di Rasch, introducendo un livello d'incertezza, si transiti da una misurazione ordinale a una a intervalli?

Abbiamo discusso due tipi di critiche/confutazioni al paradosso. Quelle del primo tipo cercano di mostrare che il modello di Guttman non sarebbe semplicemente una fattispecie del modello di Rasch "senza errore", ma che tra i due vi sarebbero differenze più profonde relative all'interpretazione delle informazioni contenute nei punteggi e, soprattutto, alla natura discreta o continua delle probabilità coinvolte. Abbiamo concluso che queste importanti osservazioni non rimuovono però, dal nostro punto di vista, la critica essenziale di Michell, ossia la mancata giustificazione dell'adozione di una funzione continua prima ancora di aver identificato il carattere più che ordinale di un attributo.

La confutazione del secondo tipo, lanciata da Zand Scholten, sottolinea invece che non sempre l'aumento dell'errore impedisce il contemporaneo miglioramento della misurazione, portando come esempio il fenomeno ondulatorio della risonanza stocastica, in cui un fattore di disturbo amplifica e rende rilevabile un segnale altrimenti molto debole. In particolare, attraverso una simulazione di dati, la studiosa mostra che lo stesso avverrebbe nel modello di Rasch, dove in determinate situazioni aumentando la discriminazione (la precisione) aumenta anche l'errore di stima dei parametri (la non affidabilità della scala). La nostra discussione dei risultati, attraverso la mediazione del concetto di "funzione informativa", ha provato a illustrare il meccanismo che porta a quegli esiti: la discriminazione di un *item* è riferita alle regioni centrali della sua curva e, se aumenta, esse progressivamente si restringono; in certe circostanze, se la distribuzione dei valori degli *item* è squilibrata rispetto a quella dei soggetti, avviene una perdita complessiva di capacità di stima del test globale, cioè un aumento dell'errore. Se gli *item* sono invece distribuiti in modo equilibrato, nella simulazione di Zand Scholten il rapporto tra errore di stima e precisione si dimostra inverso, ciò che ci si aspetterebbe in un'ottica classica e su cui si fonda la premessa dell'argomento con cui Michell individua il paradosso di Rasch. Perciò, dal nostro punto di vista, più che confermare l'analogia con la risonanza stocastica (un fenomeno che si può comunque produrre sperimentalmente, sotto il controllo dei ricercatori), l'esempio dimostra l'impatto che lo strumento di rilevazione (il test) sprigiona sulla misurazione e l'interazione che crea con la distribuzione del campione se è costruito in modo squilibrato, producendo stime inaffidabili (poco in-

formative). E' quello che, abbiamo immaginato, accadrebbe con un metro di legno che abbia una graduazione molto sensibile (sempre più sensibile) in alcuni segmenti ma sia assolutamente liscio in altri (sempre più ampi). Non ci pare, quindi, che gli argomenti di confutazione esaminati riescano a eludere in modo definitivo e convincente i termini del paradosso.

Non convince affatto, invece, l'analogia con la meccanica quantistica, utilizzata per affermare il legame tra precisione e probabilità. Il parallelo tra la fisica contemporanea e i modelli psicometrici probabilistici si ferma, per l'appunto, a un qualche utilizzo della probabilità, ma radicalmente diverso è il quadro epistemologico e il percorso che conduce alle rispettive soluzioni. Forse, più fecondi per le scienze umane, rispetto alla loro condizione attuale, sono concetti quali "indeterminazione" e "complementarità", che riconoscono il ruolo dell'apparato di rilevazione nei processi di misurazione e la coesistenza d'immagini distinte scaturite dalla definizione e dalla rilevazione dello stesso oggetto di ricerca.

Modelli formali, strutture empiriche, misurazione: problemi aperti

Nel corso del testo abbiamo necessariamente concentrato la nostra attenzione su un tema specifico, scegliendo di restringere l'attenzione sul modello di Rasch nella sua forma base dicotomica. La scelta è stata dettata da motivi precisi, primo fra tutti il fatto che il dibattito che affronta congiuntamente i temi della misurazione, dei modelli IRT e dell'ACM individua prevalentemente il modello di Rasch come fattispecie su cui far ruotare la riflessione: in pratica, è molto frequente che esso sia scelto come termine di paragone per parlare di misurazione in senso stretto e modelli probabilistici, come una notevole letteratura dimostra (parte della quale citata lungo tutto il nostro testo). In effetti, e questo è almeno il secondo motivo in gioco, per molti aspetti il modello di Rasch offre caratteristiche formali e potenzialità molto importanti per chi ambisce a sviluppare, nelle scienze psicologiche e sociali, procedure per scalare atteggiamenti e dimensioni valoriali che si avvicinino a uno standard di misurazione formalmente molto più rigoroso di altri modelli tradizionali. La sua semplicità e il suo rigore costituiscono aspetti fa-

vorevoli al controllo accurato dei requisiti di scalabilità cumulativa e, almeno da un punto di vista formale, di additività.

Ciò non deve distrarre dal fatto che, nell'ambito della IRT e talvolta a partire dallo stesso modello di Rasch, sono stati sviluppati modelli per *item* politomici, come il *Partial Credit Model* e il *Rating Scale Model*. Sarebbe interessante, giunti alla fine di questa disamina, proseguire e ampliare anche a queste fattispecie una riflessione epistemologica approfondita sugli assunti, le proprietà dei modelli e le loro connessioni con le teorie degli attributi e le tecniche di rilevazione, a partire dalla letteratura presente sul tema. Il motivo di tale interesse, e di tale taglio di ricerca, nasce dalla constatazione che il principale punto di debolezza nell'applicazione di modelli come quello di Rasch, che necessita approfondimento e indagine, non risiede al livello delle strutture formali, che hanno la configurazione adatta a elevare il rigore dei procedimenti di rappresentazione nelle scienze umane; l'anello debole risiede nel rapporto (e all'incrocio) tra l'analisi dei caratteri delle strutture empiriche, lo sviluppo delle teorie degli attributi e la costruzione degli strumenti di rilevazione: ossia nella carne viva della ricerca empirica.

Linearità della scala, statistiche sufficienti, invarianza, oggettività specifica sono caratteristiche fondamentali. Ciò richiede che il modello possa adattarsi in modo soddisfacente ai dati prodotti dalla base empirica. E che, per uno stesso attributo, ciò costituisca un risultato non sporadico, che individui una regolarità nel rapporto di rappresentazione tra l'attributo e la struttura del modello. Al contrario, l'obiettivo della confrontabilità e della generalizzabilità perde buona parte della sua plausibilità. Questo è un compito che solo l'indagine empirica può condurre, contribuendo in questo modo a irrobustire la definizione teorica degli attributi e il rapporto tra questi e gli indicatori operativi, costituiti dagli *item*.

Dal punto di vista formale e matematico del modello di Rasch, è possibile spingersi molto avanti, trovando nella teoria della misurazione additiva congiunta un'altra giustificazione assiomatica del suo carattere di misurazione a intervalli. Le critiche che abbiamo analizzato nascono, tutte, nel momento in cui dal modello si passa alla discussione del rapporto tra struttura empirica e formale. I motivi che abbiamo incontrato spaziano dagli aspetti epistemologici a quelli tecnici. In generale, si può concordare che, alla luce di alcuni degli argomenti a sfavore proposti da più parti, non è ancora possibile accettare il modello di Rasch come una completa alternativa al controllo degli assiomi

dell'ACM. I contro-argomenti elaborati non ci sembrano, sulla base della nostra discussione, risolvere del tutto le contraddizioni e i limiti: ragioni logiche, strutturali e teoriche impediscono di compiere l'ultimo passo. Ciò non implica, beninteso, che le affinità siano irrilevanti, soprattutto se potessero essere rimosse le riserve epistemologiche citate (e sempre che ciò sia possibile).

L'*input* dato da diversi studiosi è che le teorie sugli attributi (e quindi la conoscenza della loro struttura) sarebbero generalmente insufficienti a definire in modo altamente specifico le dimensioni latenti e a suggerire procedure formali rigorose per la selezione degli *item* in grado di garantire una corretta denotazione del costrutto. La pratica di ridefinire in corso d'opera il *pool* di *item*, in sede di misurazione, eliminandone alcuni per aumentare l'adattamento del modello ai dati, non è un modo per risolvere il problema ed essere più "aderenti alla realtà", ma rischia piuttosto di evocare illusorie dimensioni quantitative, senza aumentare la nostra conoscenza degli attributi. La questione rimane aperta, tutt'altro che inedita ma non meno attuale: più le teorie sono precise e riescono a produrre ipotesi articolate e complete e descrizioni accurate dei costrutti, più i test diventano strumenti capaci di controllare in modo rigoroso le ipotesi avanzate; al contempo, risultati ottenuti attraverso procedure il più possibile formalizzate, riflettute e rigorose sono l'unico modo per accrescere l'accuratezza teorica dei nostri concetti di proprietà. Questo filone di riflessione, applicato a modelli come quello di Rasch, è tutt'altro che esaurito e potrebbe aiutare nello sviluppo dei costrutti nelle nostre discipline.

Nel complesso teorico-pratico dei processi di misurazione, troviamo una dimensione legata ai vincoli imposti dalla struttura formale dei numeri reali su cui rappresentare la realtà (dimensione che per molto tempo, nella concezione della misurazione ereditata da Stevens, è stata preponderante) e una dimensione legata ai vincoli imposti dalla struttura empirica (dimensione valorizzata soprattutto dalla visione realista della teoria di N.R. Campbell): dimensioni che, nei successivi sviluppi assiomatici della RMT, trovano entrambe un certo riconoscimento (seppure autori realisti come Michell evidenzino differenze filosofiche di rilievo). Ma tra queste due dimensioni si colloca la questione della costruzione operativa della misura, ossia le procedure di rilevazione. Gli indicatori delle proprietà sono necessariamente *item* discreti e, già nel processo di rispo-

sta, intervengono in modo costitutivo i caratteri concernenti la loro chiusura, che non sono teoricamente neutri. Questo è un ulteriore punto su cui è ancora necessario soffermare l'attenzione in ricerche future. Gli stessi esiti di una domanda possono essere interpretati e riclassificati in modi diversi. Per esempio, il carattere dicotomico dei dati richiesti dal modello di Rasch permette di utilizzare dati ottenuti anche con risposte aperte, multiple ecc., successivamente ricondotte a chiusura binaria. Ma cosa cambia se la domanda è posta direttamente in modo dicotomico o ricondotta a tale esito a partire da modalità diverse di chiusura? Due campioni composti di soggetti del tutto confrontabili, dati gli stessi *item* ma con chiusure di risposta diverse, produrrebbero esiti riconducibili a risultati analoghi, oppure il processo di rilevazione intaccherebbe in qualche modo l'indipendenza dei risultati? Tale questione è tutt'altro che meramente tecnica, perché dalla risposta alle due domande precedenti passa la differenza tra la pretesa di misurare *quello stesso attributo*, in modo *indipendente dalle circostanze contingenti*, e quella di rilevare invece un costrutto frutto di *determinate circostanze*, relativo a *quel campione* di soggetti, a quegli *item* e a quella *specifica chiusura* delle risposte, caratterizzata da meccanismi propri di elaborazione cognitiva. Questo è un aspetto scarsamente approfondito nel dibattito che abbiamo incontrato trattando il tema del fondamento della misurazione nel modello di Rasch. Eppure il tipo di strumento utilizzato per la rilevazione è assolutamente decisivo e, come abbiamo commentato durante il testo, la letteratura è ricca di esempi che dimostrano come il rapporto tra proprietà, *item* e tecnica di rilevazione sia assolutamente denso e decisivo. L'analisi dei meccanismi di risposta gioca, quindi, un ruolo assolutamente importante, da approfondire e sviluppare ulteriormente e in modo adeguato, all'interno del dibattito stesso sulla misurazione e in connessione con tutti gli altri elementi, non dando per scontato che i punteggi su cui si fondano le stime siano ottenuti in modo non problematico come indicatori validi della proprietà.

Stesso discorso può essere esteso ad altri elementi, come l'ordine di presentazione, i differenziali di significato e tutto quello che ha a che vedere con la costruzione del dato e, quindi, con la fondatezza degli assunti di indipendenza e con l'oggettività del modello.

Possedere modelli matematici in grado di costruire scale a intervalli, come abbiamo riconosciuto essere potenzialmente il modello di Rasch, facilita questo compito perché garantisce una copertura robusta sul versante formale e permette di indagare tutto il

prisma di questioni epistemologiche connesse: è un primo passo verso una fondazione più rigorosa delle stime nello studio degli atteggiamenti (auspicabile, almeno per qualcuno, ma non per questo ovvia, possibile a prescindere o tantomeno obbligata). A ciò si aggiunga che, laddove si voglia rivendicare qualcosa di più del semplice ordinamento di *item* e soggetti, dovrebbe diventare buona pratica quella di utilizzare gli strumenti logici e formali a disposizione, compresi quelli non parametrici, per testare gli assunti riguardanti gli attributi, sia sulle stime sia direttamente sulla base empirica, accumulando evidenze sulla natura degli attributi. Ciò, anche laddove l'adattamento dei dati al modello fosse soddisfacente e per quanto certi strumenti assiomatici pongano l'esigenza di uno sforzo nella direzione dell'apprendimento di linguaggi logico-formali non immediati.

In questo modo si può stimolare il progresso delle teorie e delle descrizioni delle dimensioni latenti che vorremmo indagare e conoscere, innescando un circolo virtuoso. Diversamente, possedere modelli che raggiungano un livello di precisione nella "misurazione" del tutto sproporzionato rispetto alla capacità dei ricercatori di giustificare i loro assunti e alla capacità dei loro concetti di interpretare in modo significativo i risultati, riconnettendoli a teorie più generali, non solo può essere molto meno utile di ciò che sembri a prima vista, ma può anche indebitamente rallentare il successivo sviluppo dei concetti e della conoscenza del campo d'indagine.

Ringraziamenti

E’ doveroso, alla fine di questo percorso di studi, ringraziare tutti coloro i quali sono stati fonte d’ispirazione e di aiuto per portare a termine questo mio lavoro, nel quale lacune, debolezze o errori sono da attribuire solamente a chi scrive.

Ringrazio innanzitutto il Prof. Sandro Landucci, tutor della tesi, punto di riferimento non solo disciplinare, ma soprattutto umano. Il suo costante incoraggiamento e il suo appoggio, dal primo all’ultimo giorno, non saranno da me mai dimenticati.

I ringraziamenti più sentiti e la mia riconoscenza vanno al Prof. Alessandro Bruschi, fonte inesauribile di suggerimenti, che con disponibilità e pazienza ha seguito la stesura del testo in tutte le sue fasi, offrendo un supporto assolutamente indispensabile al mio lavoro con giudizi, valutazioni e correzioni, onorandomi del suo tempo e della sua fiducia.

Un ringraziamento grandissimo va anche alla prof.ssa Filomena Maggino, la cui competenza e i cui commenti sono stati una base fondamentale per sviluppare la mia riflessione, soprattutto per quanto riguarda la parte tecnica e statistica e, non meno importante, per stimolare il mio senso critico: i nostri confronti sono stati una fonte di riflessione e interesse.

Un grazie di cuore anche alla dott.ssa Ivana Acocella, per la sua impareggiabile opera d’orientamento in questi anni e la rassicurante e amichevole presenza. E per avermi aperto le porte del suo corso di Ricerca qualitativa.

Un ringraziamento al prof. Marco Bontempi, coordinatore della Scuola di dottorato, per il lavoro svolto a supporto di noi studenti.

Al prof. Alberto Marradi va qualcosa di più e di diverso di un ringraziamento: il riconoscimento di un debito intellettuale profondo, del merito (o della colpa) di avermi fatto conoscere e amare la metodologia delle scienze sociali, e l’ammirazione che sempre un modesto allievo deve al proprio Maestro.

Un grazie va a tutti gli altri docenti che in questi anni mi hanno concesso di seguire i loro corsi e seminari, permettendomi di colmare lacune, sollevare dubbi e suscitare idee per me preziose. In particolare, citando solo alcuni ma pensando anche a tutti gli altri, la mia gratitudine va alla prof.ssa Erika Cellini, alla prof.ssa Angela Perulli e al prof. Filippo Buccarelli, con i quali le strade si sono regolarmente intrecciate nei miei dieci anni da studente all'Università di Firenze e che tanto hanno avuto da insegnarmi; al prof. Alessandro Pagnini, le cui lezioni di Filosofia della scienza sono state un'esperienza indelebile; al prof. Pietro Luigi Clemente e alle dott.sse Caterina di Pasquale e Costanza Lanzara: il loro corso di Metodologia della ricerca demo-etno-antropologica è stata una vera e propria, lunga immersione nella passione per la ricerca e per lo studio degli esseri umani; ai proff. Giancarlo Rovati e Marco Caselli, per aver accolto con gentilezza e rispetto un "forestiero" in un loro ciclo di seminari presso l'Università Cattolica di Milano.

Infine, un ringraziamento va ai miei colleghi di dottorato e agli altri docenti, ricercatori e assegnisti del Dipartimento di Scienze Politiche e Sociali dell'Università di Firenze.

Non potrei terminare senza ringraziare tutti i miei cari e i miei amici, a cominciare dalla mia amata compagna Clara e da mia madre: senza queste persone, niente sarebbe possibile e niente avrebbe senso.

BIBLIOGRAFIA

ANDERSEN, Erling B. [1973] *A Goodness of Fit Test for the Rasch Model*, in "Psychometrika", vol. XXXVIII, n.1, pp. 123-40.
- [1977] *Sufficient Statistics and Latent Trait Models*, in "Psychometrika", vol. XLII, n.1, pp. 69-81.

ANDRICH, David [1978] *A rating formulation for ordered response categories*, in "Psychometrika", vol. XLIII, n. 4, pp. 561-73.
- [1988] *Rasch models for measurement*, Newbury Park, SAGE.
- [2003] *On the distribution of measurements in units that are not arbitrary*, in "Epistemology of Measurement", vol. XLII, n. 4, pp. 557-89.

BAKER, Frank B. [2001] *The Basics of Item Response Theory*, ERIC.

BAKER, Frank B. e Seock-Ho KIM [2004] *Item Response Theory*, New York, Dekker.

BARBARANELLI, Claudio e Elena NATALI [2005] *I test psicologici: Teorie e modelli psicometrici*, Roma, Carocci.

BARRET, Paul [2003] *Beyond psychometrics. Measurement, non-quantitative structure, and applied numerics*, in "Journal of Managerial Psychology", vol. XVIII, n. 5, pp. 421-439.
- [2008] *The Consequence of Sustaining a Pathology: Scientific Stagnation*, in "Measurement", vol VI, n. 1, pp. 78-123.

BERGSON, Henri [1889] *Essai sur les données immédiates de la conscience*; trad. ing. *Time and Free Will: An Essay on the Immediate Data of Consciousness*, New York, Dover Publications, 2001.

BIRNBAUM, A [1968] *Some Latent Trait Models and Their Use in Inferring an Examinee's Ability*, in F. M. Lord e M. R. Novick (a cura di) *Statistical Theories of Mental Test Scores*, Reading, Addinson-Wesley.

BOGARDUS, Emory S. [1925], *Measuring social distances*, in M. Fishbein (a cura di), *Attitude theory and measurement*, New York, John Wiley, 1967.

BONCORI, Lucia [2006] *I test in psicologia*, Bologna, il Mulino.

BORING, Edwin G. [1920] *The logic of the normal law of error in mental measurement*, in "The American Journal of Psychology", vol. XXXI, n. 1, pp. 1-33.

BORSBOOM, Denny [2005] *Measuring the mind: Conceptual issues in contemporary psychometrics*, Cambridge, University Press.

BORSBOOM, Denny e Gideon J. MELLENBERGH [2004] *Why Psychometrics is Not Pathological: A Comment on Michell*, in "Theory & Psychology", vol. XIV, n. 1, pp. 105-120.

BORSBOOM, Denny e Annemarie ZAND SCHOLTEN [2008] *The Rasch Model and Conjoint Measurement Theory from the Perspective of Psychometrics*, in "Theory & Psychology", vol. XVIII, n. 1, pp. 111-7.

BRIDGMAN, Percy W. [1927] *The Logic of Modern Physics*; trad. it. *La logica della fisica moderna*, Torino, Boringhieri, 1961.

BROGDEN, Hubert E. [1977] *The Rasch model, the lay of comparative judgement and additive conjoint measurement*, in "Psychometrika", vol. LXII, n. 4, pp. 631-4.

BRUSCHI, Alessandro [1999] *Metodologia delle scienze sociali*, Milano, Bruno Mondadori.

- [2005] *Metodologia della ricerca sociale*, Bari, Laterza.

BURRO, Roberto [2007] *La misurazione fondamentale*, in "Dipav Quaderni", n. 20, pp.

9-24.
- [2009] *La teoria della misurazione congiunta in psicofisica*, in "Dipav Quaderni", n. 25, pp. 119-30.

CAMPBELL, Donald T. e Donald W. FISKE [1959] *Convergent and discriminant validation by the multitrait-multidimethod matrix"*, in "Psychological Bullettin", vol. LVI, n. 2, pp. 81-105.

CAMPBELL, Norman R. [1920] *Physics: The elements*, Cambridge, University Press.

CANNAVÒ, Leonardo [1999] *Teorie e problemi della misurazione sociale*, Milano, LED.
- [2003] *Oltre Thurstone e Likert. La valutazione di atteggiamenti e motivazioni con la tecnica TLL*, Roma, La Goliardica.

CLIFF, Norman [1992] *Abstract Measurement Theory and the Revolution That Never Happened*, in "Psychological Science, vol. III, n. 3, pp. 186-90.

COHEN, Morris R. e Ernst NAGEL [1934] *An Introduction to Logic and Scientific Method*, New York, Harcourt.

COOMBS, Clyde H. [1950] *Psychological Scaling Without a Unit of Measurement*, in "Psychological Review", vol. LVII, n. 3, pp. 145-58.
- [1964] *A Theory of Data*, New York, Wiley.

CORBETTA, Piergiorgio [2003] *La ricerca sociale: metodologia e tecniche. Volume 2. Le tecniche quantitative*, Bologna, il Mulino.

CRISTANTE, Francesca e Stefania MANNARINI [2004] *Misurare in psicologia. Il modello di Rasch*, Bari, Laterza.

de AYALA, Rafael J. [2009] *The Theory and Practice of Item Response Theory*, New York, The Guilford Press.

DIANA, Paolo [2008] *La rilevazione degli atteggiamenti e delle opinioni: il contributo di Louis Leon Thurstone*, Lecce, Kurumuny; introduzione a Louis L. Thurstone, *Attitude Can Be Measured*, 1928, trad. it.

DUHEM, Pierre [1914] *The aim and structure of physical theory*, trad. inglese, Princeton, Princeton University Press, 1954.

EDWARDS, Allen L. [1957] *The Social Desiderability Variable in Personality Assessment and Research*, New York, Dryden.

EMBRETSON, Susan E. e Steven P. REISE [2000] *Item Response Theory for Psychologists*, Mahwah, Lawrence Erlbaum Associates.

FALMAGNE, Jean-Claude [1976] *Random conjoint measurement and loudness summation*, in "Psychological Review", vol. LXXXIII, pp. 65-84.

FANELLI, Daniele [2009] *How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data*, in "PLoS ONE", vol. IV, n. 5, e5738, doi:10.1371/journal.pone.0005738.

- [2010] *"Positive" Results Increase Down the Hierarchy of the Sciences*, in "PLoS ONE", vol. IV, n. 4, e10068, doi:10.1371/journal.pone.0010068.

FERGUSON, Andrew *et al.* [1940] *Quantitative estimates of sensory events: final report*, in "Advancement of Science", vol. I, pp. 331-49.

FISCHER, Gerard H. [1968] *Psychologische Testtheorie*, Berna, Huber.

- [1974] *Einführung in die Theorie psychologischer Tests*, Berne, Huber.

- [1995a] *Some Neglected Problems in IRT*, in "Psychometrika", vol. LX, n. 4, pp. 459-87.
- [1995b] *Derivations of the Rasch model*; in Gerard H. Fischer e Ivo W. Molenaar (a cura di), *Rasch models: Foundations, recent developments, and applications*, New York, Springer.

FROVA, Andrea e Mariapiera MARENZANA [1998] *Parola di Galileo*, Milano, Rizzoli.

FRUDA', Luigi [1975] *Elementi di metodologia e tecnica della ricerca sociale*, Roma, Elia.

GAMMAITONI, Luca, Peter HÄNGGI, Peter JUNG, Fabio MARCHESONI [1998] *Stochastic resonance*, in "Reviews of Modern Physics", vol. LXX, n. 1, pp. 223-87.

GIAMPAGLIA, Giuseppe [1990] *Lo scaling unidimensionale nella ricerca sociale*, Napoli, Liguori.
- [2008] *Il modello di Rasch nella ricerca sociale*, Napoli, Liguori.

GIGLIOLI, Pier Paolo e Giolo FELE (a cura di) [2000] *Linguaggio e contesto sociale*, Bologna, il Mulino.

GOODENOUGH, Ward H. [1944] *A Technique for Scale Analysis*, in "Educational and Psychological Measurement", vol. IV, n. 3, pp. 179-80.

GREEN, Bert F. [1954] *Attitude Measurement*, in G. Lindzey (a cura di) *Handbook of Social Psychology, Volume I*, Reading, Addison-Wesley.
- [1956] *A method of scalogram analysis using summary statistics*, in "Psychometrika", vol. XXI, pp. 79-88.

GREEN, Kathy E. [1986] *Fundamental Measurement: A Review and Application of*

Additive Conjoint Measurement in Educational Testing, in "The Journal of Experimental Education", vol. LIV, n. 3, pp. 141-7.

GUTTMAN, Louis A. [1944] *A Basis for Scaling Qualitative Data*, in "American Sociological Review", vol. IX, pp. 139-50.
- [1950] *The basis for scalogram analysis*, in Samuel Stouffer et al. *Measurement and Prediction*, The American Soldier, vol. IV., New York, Wiley.

HACKING, I. [1965] *The logic of statistical inference*, Cambridge, University Press.

HAMBLETON, Ronald K. [1994] *Item Response Theory: a Broad Psychometric Framework for Measurement Advances*, in "Psicothema", vol. VI, n.3, pp. 535-56.

HAMBLETON, Ronald K., Hariharan SWAMINATHAN, H. Jane ROGERS [1991] *Fundamentals of item response theory. Measurement Methods for the Social Sciences series*, vol. II, Londra, SAGE Publications.

HEENE, Moritz [2013] *Additive conjoint measurement and the resistance toward falsifiability in psychology*, in "Frontiers in Psychology", vol. IV, pp. 246.

HEISENBERG, Werner [1955/1957] *Natura e fisica moderna*, trad. it., Milano, Garzanti, 1957.

von HELMOLTZ, Hermann [1887] *Zählen und Messen erkenntnistheoretisch betrachtet*, in *Philosophische Aufsatze, Eduard Zeller zu seinem funfzigjahrigen Doctorjubilaum gewidmet*; trad. ing. *Numbering and measuring from an epistemological viewpoint*, in *Epistemological writings*, 1977.

HUGHES, John A. e Wesley W. SHARROCK [1997] *The Philosophy of Social Research*; trad. it. *Filosofia della ricerca sociale*, Bologna, il Mulino, 2005.

JÖRESKOG, Karl G. [1969] *A general approach to confirmatory maximum likelihood factor analysis*, in "Psychometrika", vol. XXXIV, pp. 183-202.

KARABATSOS, George [2001] *The Rasch Model, Additive Conjoint Measurement, and New Models of Probabilistic Measurement Theory*, in "Journal of Applied Measurement", vol. II, n.4, pp. 389-423.

KEATS, John A. [1967] *Test theory*, in "Annual Review of Psychology", vol. XVIII, pp. 217-38.

KERLINGER, Fred N. e Howard B. LEE [2000] *Foundations of behavioral research*, Orlando, Harcourt College Publishers.

KLINE, Paul [2000] *A psychometrics primer*, Londra, Free Association Books.

KRANTZ, David [1964] *Conjoint measurement: the Luce — Tukey axiomatisation and some extensions*, in "Journal of Mathematical Psychology", vol. I, n. 2, pp. 248-77.

KRANTZ, David H., R. Duncan LUCE, Patrick SUPPES e Amos TVERSKY [1971] *Foundation of Measurement, Volume I: Additive and Polynomial Representations*, Mineola, Dover Publication.

KUHN, Thomas [1970] *The structure of Scientific Revolutions*, seconda edizione; trad. it. *La struttura delle rivoluzioni scientifiche*, Torino, Einaudi.

KYNGDON, Andrew [2008a] *The Rasch Model from the Perspective of the Representational Theory of Measurement*, in "Theory Psychology", vol. XVIII, n. 1, pp. 89-109.

[2008b] *Conjoint Measurement, Error and the Rasch Model: A Reply to Michell, and Borsboom and Zand Scholten*, in "Theory Psychology", vol. XVIII, n. 1, pp. 125-31.

- [2011] *Plausible measurement analogies to some*

psychometric models of test performance, in "British Journal of Mathematical and Statistical Psychology", vol. LXIV, pp. 478–97.

LA SALA, Diego [2010] *Giovani e valori, una ricerca su legalità, partecipazione ed ambiente tra gli studenti universitari fiorentini: i risultati e alcuni contributi metodologici*, Università di Firenze, tesi di laurea magistrale.

LEHMAN, Richard S. [1991] *Statistics and research design in the behavioral sciences*, Belmont, Wadsworth/Thomson Learning.

LORD, Frederic M. [1980] *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale, Lawrence Erlbaum.

LORD, Frederic M. e Melvin R. NOVICK [1968] *Statistical Theories of Mental Test Scores*, Reading, Addinson-Wesley.

LUCE, R. Duncan [1987] *Measurement structures with Archimedean ordered translation groups*, in "Order", vol. IV, pp. 165-89.

LUCE, R. Duncan, David H. KRANTZ, Patrick SUPPES e Amos TVERSKY [1990] *Foundations of Measurement. Volume 3. Representation, Axiomatization, and Invariance*, San Diego, Academic Press.

LUCE, R. Duncan e Louis Narens [1994] *Fifteen problems concerning the representational theory of measurement*, in Patrick Humphreys (a cura di), *Patrick Suppes: Scientific philosopher*, vol. II, pp. 219-49, Dordrecht, Kluwer Academic.

LUCE, R. Duncan e John. W. TUKEY [1964] *Simultaneous conjoint measurement: a new scale type of fundamental measurement*, in "Journal of Mathematical Psychology", vol. I, pp. 1–27.

MAGGINO, Filomena [2004a] *La misurazione nella ricerca sociale*, Firenze, University Press.
- [2004b] *I modelli di scaling*, Firenze, University Press.
- [2007] *La rilevazione e l'analisi statistica del dato soggettivo*, Firenze, University Press.

MARRADI, Alberto [1980] *Concetti e metodo per la ricerca sociale*, Firenze, Giuntina.
- [1981] *Misurazione e scale: qualche riflessione e una proposta*, in "Quaderni di Sociologia", vol. XXIX, n. 4, pp. 595-639.
- [1985] *Unità di misura e unità di conto*, in "Rassegna Italiana di Sociologia", vol. XXVI, n. 2, pp. 229-38.
- [2002] *Linee guida per l'analisi bivariata dei dati nelle scienze sociali*, Milano, Franco Angeli.
- [2005] *Raccontar storie*, Roma, Carocci.
- [2007] *Metodologia delle scienze sociali*, Bologna, il Mulino.

MASTERS, Geofferey N. [1982] *A Rasch model for partial credit scoring*, in "Psychomerika", vol. XLVII, n. 2, pp. 149-74.

MCIVER, John P. e Edward G. CARMINES [1981] *Unidimensional Scaling*, Newbury Park, Sage.

MCNAMARA, Bruce e Kurt WIESENFELD [1989] Theory of stochastic resonance , in "Physical Review A", vol. XXXIX, n. 9, pp. 4854-69.

MENZEL, Herbert [1953] *A new coefficient for scalogram analysis*, in "Public Opinion Quarterly", vol. XVII, pp. 268-80.

MICHELL, Joel [1988] *Some problems in testing the double cancellation condition in conjoint measurement*, in "Journal of Mathematical Psychology", vol. XXXII, pp. 466–73.
- [1994] *Measuring dimensions of belief by unidimensional unfolding*,

in “Journal of Mathematical Psychology”, vol. XXXVIII, pp. 244-73.
- [1997] *Quantitative science and the definition of measurement in psychology*, in “British Journal of Pyhology”, vol. LXXXVIII, pp. 355-83.
- [2000] *Normal Science, Pathological Science and Psychometrics*, in “Theory & Psychology”, vol. X, n. 5, pp. 639-67.
- [2004] *Item Response Models, Pathological Science and the Shape of Error*, in “Theory & Psychology”, vol. XIV, n. 1, pp. 121-9.
- [2007] *Measurement*, in S. P. Turner e M. W. Risjord (a cura di), *Handbook of the Philosophy of Science. Philosophy of Anthropology and Sociology*, Amsterdam, North Holland.
- [2008a] *Is Psychometrics Pathological Science?*, in “Measurement”, vol. VI, n. 1-2, pp. 7-24.
- [2008b] *Conjoint Measurement and the Rasch Paradox: A Response to Kyngdon*, in “Theory Psychology”, vol. XVIII, n. 1, pp. 119-24.
- [2009] *The psychometricians’ fallacy: Too clever by half?*, in “British Journal of Mathematical and Statistical Psychology”, vol. LXII, pp. 41–55.

MOKKEN, Robert J. [1971] *A theory and procedure of scale analysis*, The Hague, Mouton.

MONTUSCHI, Eleonora [2006] *Oggettività e scienze umane*, Roma, Carocci.

MORETTO, Antonio [2012] *La teoria elementare della misura delle grandezze da un punto di vista filosofico-matematico*, in P. Di Nicola (a cura di), *La sfida della misurazione nelle scienze sociali*, Milano, Franco Angeli.

NICKERSON, Carol A. e Gary H. MCCLELLAND [1984] *Scaling distortion in numerical conjoint measurement*, in “Applied Psychological Measurement”, vol. VIII, n.2, pp. 183-98.

NOVENTA, Stefano e Giulio VIDOTTO [2012a] *Da Fechner a Luce, la misura in psicologia – Parte I*, in "Tutto_Misure", vol. XIV, n.2, pp. 127-30.
[2012b] *Da Fechner a Luce, la misura in psicologia – Parte II*, in "Tutto_Misure", vol. XIV, n.3, pp. 193-6.

NUNNALLY, Jum C. [1978] *Psychometric Theory*, New York, Mc-Graw Hill.

NUNNALLY, Jum C. e Ira H. BERNSTEIN [1994] *Psychometric Theory*, New York, Mc-Graw Hill.

PEDON, Arrigo [1995] *Metodologia per le scienze del comportamento*, Bologna, il Mulino.

PERLINE, Richard, Benjamin D. WRIGHT e Howard WAINER [1979] *The Rasch Model as Additive Conjoint Measurement*, in "Applied Psychological Measurement", vol. III, n. 2, pp. 237-55.

PICCOLO, Domenico [2004] *Statistica per le decisioni*, Bologna, il Mulino.

PITRONE, Maria C. [2009] *Sondaggi e interviste*, Bologna, il Mulino.
- [2012] *Il problema della misurazione nelle scienze umane*, in P. Di Nicola (a cura di), *La sfida della misurazione nelle scienze sociali*, Milano, Franco Angeli.

POPPER, Karl R. [1959], *The propensity interpretation of probability*, in "The British Journal for the Philosophy of Science", vol. X, n. 37, pp. 25–42.

PRIMI, Caterina e Francesca CHIESI [2005] *Introduzione alla psicometria*, Roma-Bari, Laterza.

RASCH, Georg [1960] *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen, Danish Institute for Educational Research.

- [1961] *On General Laws and the Meaning of Measurement in Psychology. Proceedings of the IV Symposium of Mathematical Statistics and Probability*, vol. IV, pp. 321-33, Berkeley, University of California Press.

- [1967] *An Informal Report on a Theory of Objectivity in Comparisons*, in L J. van der Kamp e C A. Vlek (a cura di) *Measurement Theory*, University of Leyden.

- [1968] *A Mathematical Theory of Objectivity and Its Consequences for Model Construction*, European Meeting on Statistics, Econometrics, and Management Science, Amsterdam.

- [1972] *Objectivitet i samfundsvidenskaberne et metodeproblem*, in "National-økonomisk Tidsskrift", n. 110, pp. 161-96.

- [1977] *On Specific Objectivity. An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements*, in M Blegvad (a cura di) *The Danish Yearbook of Philosophy*, Copenaghen, Munksgaard.

REISE, Steven P., Andrew T. AINSWORTH e Mark G. HAVILAND [2005] *Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research*, in "Current Directions in Psychological Science", vol. XIV, n. 2, pp. 95-101.

RICOLFI, Luca [1985] *Operazioni di ricerca e scale*, in "Rassegna Italiana di Sociologia", vol. XXVI, n. 2.

ROBERT, Fred S. [1979] *Measurement Theory*, Reading, Addison-Wesley.

ROBUSTO, Egidio e Pasquale ANSELMI [2012] *La misurazione fondamentale in psicologia*, in P. Di Nicola (a cura di), *La sfida della misurazione nelle scienze sociali*, Milano, Franco Angeli.

RUBINI, Vittorio [1984] *Test e misurazioni psicologiche*, Bologna, il Mulino.

RUSSELL, Bertrand [1903] *Principles of Mathematics*, Cambridge, University Press.

SCHEIBLECHNER, Hartman [1999] *Additive Conjoint Isotonic Probabilistic Models (ADISOP)*, in "Psychometrika", vol. LXIV, n. 3, pp. 295-316.

SCOTT, Dana [1964] *Measurement structures and linear inequalities*, in "Journal of Mathematical Psychology", vol. I, n. 2, pp. 233-47.

SCOTT, William A. [1968] *Attitude Measurement*, in G. Lindzey e E. Aronson (a cura di) *Handbook of Social Psychology*, II ed., vol. II, Reading, Addison-Wesley.

SHERRY, David [2011] *Thermoscopes, thermometers, and the foundations of measurement*, in "Studies in History and Philosophy of Science", vol. XLII, pp. 509-24.

SIJTSMA, Klaas [2012] *Psychological measurement between physics and statistics*, in "Theory & Psychology", vol. XXII, n. 6, pp. 786-809.

SIMONOTTO, Enrico, Massimo RIANI, Charles SEIFE, Mark ROBERTS, Jennifer TWITTY e Frank MOSS [1997] *Visual Perception of Stochastic Resonance*, in "Physical Review Letters", vol. LXXVIII, n. 6, pp. 1186-9.

SPARTI, Davide [2002] *Epistemologia delle scienze sociali*, Bologna, il Mulino.

STERNBERG, Saul [1969] *The discovery of processing stages: Extensions of Donders' Method*, in "Acta Psychologica", vol. XXX, pp. 276-315.

STEVENS, Stanley S. [1946] *On the Theory of Scales of Measurement*; trad. it. *La teoria delle scale di misura*, in M. Cardano e R. Miceli (a cura di), *Il linguaggio delle variabili*, Torino, Rosenberg&Sellier, 1991.

- [1968] *Measurement, Statistics, and the schemapiric view*, in "Science", n. 161, pp. 849-56.

STOUFFER, Samuel A. *et al.* [1950] *Measurement and Prediction*, Princeton, University Press.

SUPPES, Patrick, David H. KRANTZ, R. Duncan LUCE e Amos TVERSKY [1989] *Foundation of Measurement, Volume II: Geometrical. Threshold, and Probabilistic Representations*, Mineola, Dover Publication.

SUPPES, Patrick e Joseph ZINNES [1963] *Basic Measurement Theory*, in R. Duncan Luce (a cura di) *Handbook of Mathematical Psychology*, Oxford, Wiley.

TESSITORE, Cristiano [2011] *Modelli IRT e Differential Item Functioning: potenzialità e prospettive per la ricerca sociale*, Università di Firenze, tesi di dottorato.

THURSTONE, Louis L. [1927] *The Method of Paired Comparison for Social Values*, in "Journal of Abnormal and Social Psychology", vol. XXI, pp. 384-97.

- [1928] *Attitudes Can Be Measured*, in "American Journal of Sociology", vol. XXXIII, n. 4, pp. 529-54.

- [1957] *Multiple Factor Analysis: a Development and Expansions of the Vector of Mind*, Chicago, University Press.

- [1959] *The Measurement of Values*, Chicago, University Press.

THURSTONE, Louis L. e Ernest J. CHAVE [1929] *The Measurement of Attitude*, Chicago, University Press.

TORGERSON, Warren S. [1958] *Theory and Methods of Scaling*, New York, Wiley.

TRENDLER, Günther [2009] *Measurement Theory, Psychology and the Revolution*

That Cannot Happen, in “Theory & Psychology”, vol. XIX, n. 5, pp. 579-99.

VAUTIER, Stéphane, Michiel VELDHUIS, Émilie LACOT e Nadine MATTON [2012] *The ambiguous utility of psychometrics for the interpretative foundation of socially relevant avatars*, in “Theory & Psychology”, vol. XXII, n. 6, pp. 810-22.

WHITELY, Bernard E. (jr.) [1996] *Principles of Research in Behavioral Science*, Mountain View, Mayfield Publishing Co.

WHITELY, Susan E. [1977] *Models, Meanings and Misunderstandings: Some Issues in Applying Rasch's Theory*, in “Journal of Educational Measurement”, vol. XIV, n. 3, pp. 227-35.

WHITELY, Susan E. e Renè V. DAWIS [1974] *The Nature of Objectivity with the Rasch Model*, in “Journal of Educational Measurement”, vol. XI, n. 3, pp. 163-78.
- [1976] *The Influence of Test Context On Item Difficulty*, in “Educational and Psychological Measurement”, n. 36, pp. 329-37.

WRIGHT, Benjamin D. [1968] *Sample-free test calibration and person measurement*, proceedings of the 1967 Invitational Conference on Testing Problems, Princeton, Educational Testing Service.
- [1977a] *Solving measurement problems with the Rasch model*, in “Journal of Educational Measurement”, vol. XIV, n. 2, pp. 97-116.
- [1977b] *Misunderstanding the Rasch Model*, in “Journal of Educational Measurement”, vol. XIV, n. 2, pp. 219-25.

WRIGHT, Benjamin D. e Geofferey N. MASTERS [1982] *Rating Scale Analysis*, Chicago, MESA Press.

WRIGHT, Benjamin D. e Mark Stone [1999] *Measurement Essentials*, Wilmington, Wide Range.

ZAND SCHOLTEN, Annemarie [2011] *Admissible statistics from a latent variable perspective*, Università di Amsterdam, tesi di dottorato.

ZOHAR, Danah e Ian MARSHALL [1994] *The Quantum Society*, Morrow, HarperCollins.