I MODELLI DI RASCH NELLA VALUTAZIONE DELLA DIDATTICA UNIVERSITARIA

Sivia Bacci

Dipartimento di Statistica "G. Parenti", Università degli Studi di Firenze - Viale Morgagni, 59 - 50134 Firenze
e-mail: s.bacci@ds.unifi.it

Riassunto

Il problema della valutazione della qualità dei servizi offerti da sistemi complessi, quali quello universitario, è l'oggetto del presente contributo. In particolare, l'interesse è volto a misurare la soddisfazione degli studenti frequentanti per la didattica universitaria. Data la natura latente del fenomeno studiato, da un punto di vista metodologico si pone il problema di individuare opportuni strumenti statistici per pervenire ad una misura oggettiva della soddisfazione, in grado di attuare una sintesi delle risposte fornite dagli studenti frequentanti ad un questionario ad hoc.

A tale scopo vengono analizzate le potenzialità dei modelli di Rasch, quale metodo di riferimento per la valutazione di sistemi complessi. L'analisi empirica è condotta sui dati raccolti presso l'Ateneo di Firenze negli anni 2003, 2004 e 2005.

1. INTRODUZIONE

Seguendo l'esempio di altri Paesi Occidentali, quali Gran Bretagna, Stati Uniti, Australia, ecc. da qualche anno anche in Italia si sta diffondendo la cultura della valutazione dei servizi pubblici, a fini non meramente conosci-

tivi o di controllo ex post, ma anche e soprattutto a fini decisionali (Gori & Vittadini (1999) e Chiandotto (2002)).

A seguito delle riforme da cui è stato investito nel corso degli ultimi anni, il sistema dell'istruzione e, in particolare, dell'istruzione universitaria (D.M. 509 del 3/11/1999, D.M. 4/8/2000 "Determinazione delle classi delle lauree universitarie", D.M. 28/11/2000 "Determinazione delle classi delle lauree specialistiche") è sicuramente il settore pubblico nel quale la valutazione deve essere sentita come un elemento prioritario (Bini & Chiandotto (2003) e Chiandotto (2004)). Tra i numerosi elementi di efficienza ed efficacia (interne e/o esterne) che possono essere presi in esame per formulare un giudizio sul sistema universitario, un aspetto di particolare interesse è rappresentato dalla soddisfazione espressa dai più immediati e diretti fruitori del sistema educativo, gli studenti frequentanti. Conoscere l'opinione degli studenti sugli insegnamenti e sul corso di laurea che stanno frequentando è sicuramente un'informazione di rilievo, benché non l'unica, che può essere utile agli organi decisionali (singolo docente, consiglio di corso di laurea, consiglio di facoltà, ecc.) per individuare elementi di inefficienza e inefficacia del sistema e, quindi, per porre in essere opportune azioni correttive.

Poiché la soddisfazione degli studenti può essere misurata soltanto in modo indiretto attraverso le risposte fornite a domande su aspetti parziali che contribuiscono a definire il concetto di soddisfazione complessiva, è di particolare importanza l'individuazione di metodi di misura atti a tradurre l'informazione derivante da questo insieme di domande osservabili (item), indicatori parziali della variabile latente, in una misura sintetica e, per quanto possibile, oggettiva della medesima. A tal proposito, i modelli di Rasch costituiscono certamente un adeguato contesto di riferimento, date le proprietà di cui essi godono, quali l'unidimensionalità, la sufficienza dei punteggi grezzi degli item e degli individui, l'indipendenza locale degli item, la specifica oggettività, tali da garantire la traduzione nell'ambito delle

scienze sociali del concetto di misura proprio delle scienze fisiche.

Oggetto del presente contributo è una verifica delle potenzialità dei modelli di Rasch quale metodo di riferimento per la valutazione delle performance di un sistema complesso; in particolare, l'attenzione è rivolta alla valutazione del sistema universitario, in termini di soddisfazione degli studenti frequentanti.

Nel secondo paragrafo vengono analizzate le caratteristiche dei modelli di Rasch e la loro utilità a fini di valutazione, mentre nel terzo paragrafo, dopo una breve descrizione dei dati impiegati relativi agli studenti frequentanti dell'Ateneo di Firenze negli anni 2003, 2004 e 2005, si procede ad una descrizione dei risultati conseguiti attraverso l'analisi empirica effettuata. Il lavoro si conclude con alcune considerazioni sui possibili sviluppi, volte soprattutto a tenere conto in modo esplicito della struttura gerarchica dei dati (studenti in insegnamenti, insegnamenti in corsi di laurea, ecc.).

2. I MODELLI DI RASCH

2.1 IL CONCETTO DI MISURA

Misurare un oggetto rispetto ad una determinata variabile significa collocare l'oggetto stesso lungo un continuum numerico immaginario (con unità di misura convenzionale), in modo tale che sia possibile esprimere un giudizio quantitativo sulla posizione occupata sia in termini assoluti che relativi. A tale proposito, i questionari rappresentano un'utile descrizione del profilo *qualitativo* del fenomeno studiato, ma la somma dei punteggi conseguiti nelle diverse domande non ha alcun significato di tipo *quantitativo*: a titolo di esempio, se un individuo afferma che ritiene adeguati i locali in cui si svolgono le lezioni (aule=1) mentre un altro li ritiene inadeguati (aule=0), si può soltanto affermare che il primo è più soddisfatto del secondo, ma non si è in grado di quantificare questo maggior livello di soddisfazione. In altri termini, i valori numerici assegnati alle possibili risposte alle domande di un questionario (0/1 piuttosto che 0/10 oppure 0/1/2/3 piuttosto che

1/3/5/7, ecc.) hanno una natura arbitraria e, quindi, definiscono una scala qualitativa ordinale. Ciò che invece caratterizza il concetto di misura è la possibilità di collocare i conteggi osservati su una scala quantitativa.

In concreto, una misura per definizione deve possedere due requisiti imprenscindibili, tra loro fortemente correlati: **uni-dimensionalità** e **specifica oggettività**. L'uni-dimensionalità si riferisce al fatto che l'attributo latente rispetto al quale viene eseguita la misura è soltanto uno: ciò significa che gli item di un questionario sono indicatori parziali della medesima variabile latente; se questo non si verifica, dovranno previamente essere individuati sottoinsiemi omogenei di item e condurre analisi separate per ciascuno di essi. La specifica oggettività si riferisce, invece, al fatto che il processo di misurazione non deve essere influenzato da caratteristiche dell'individuo diverse da quella d'interesse oppure da altri individui o dalle peculiarità dello strumento (questionario) impiegato (Gori, Sanarico & Plazzi 2005). In altri termini, se il soggetto A è più soddisfatto del soggetto B relativamente ad un certo insegnamento, tale relazione deve rimanere invariata al modificarsi delle caratteristiche individuali e del questionario impiegato (a meno che questionari diversi misurino aspetti diversi della soddisfazione complessiva, quali ad es. soddisfazione per gli aspetti organizzativi e soddisfazione per gli argomenti trattati: ma in tal caso viene perso il requisito della uni-dimensionalità).

Un valido modello di misura deve, dunque, essere tale da garantire il rispetto della uni-dimensionalità e della specifica oggettività; a tal proposito, la Item Response Theory - IRT (Baker & Kim 2004) risolve il problema della traduzione dei conteggi discreti osservabili in manifestazioni di un continuum latente tramite il ricorso ad un contesto probabilistico: il punteggio grezzo di "1" ad un item, ad esempio, viene trasformato nella probabilità attesa - quindi in un valore compreso nell'intervallo continuo [0, 1] - di osservare una risposta pari ad "1" (e in un termine di errore). Un'ulteriore

trasformazione in logit[1] - quindi in un valore compreso nell'intero asse reale - si rivela poi necessaria per evitare che a soggetti con livelli estremi, ma differenti, del tratto latente sia assegnata la medesima probabilità.

Tra i vari modelli IRT, il modello di Rasch è forse l'unico (Gori et al. 2005) in grado di garantire il rispetto dei requisiti della misura, in virtù delle proprietà di cui gode.

2.2 IPOTESI E STRUTTURA DEI MODELLI DI RASCH

Partendo dalle risposte osservate, il modello di Rasch si pone l'obiettivo di spiegare come varia la probabilità di osservare un certo *pattern* di risposte in funzione del tratto latente misurato. Questa probabilità dipende da due soli tipi di parametri: i parametri β_j $(j = 1, 2, ..., J)$ di "difficoltà" dei J item e i parametri di "abilità" θ_i $(i = 1, 2, ..., I)$ degli I individui[2]. Sotto l'ipotesi di item dicotomici, il primo tipo di parametri indica il livello di criticità di ciascuna domanda del questionario: quanto più un item è difficile e tanto meno è probabile osservare individui che scelgono la modalità di risposta "1" piuttosto che "0". Il secondo tipo di parametri, invece, fa riferimento al livello in cui il tratto latente di interesse è presente in ciascun soggetto, indica cioè la misura della variabile latente corrispondente ad un determinato punteggio conseguito sul questionario.

La relazione matematica che lega i due parametri è resa esplicita dal

---

[1] *Oltre al modello logistico, in letteratura sono previsti anche altri modelli probabilistici, quale quello normale, che sono però decisamente meno ricorrenti nelle applicazioni pratiche, perché a fronte di una maggiore complessità matematica (il modello normale prevede un integrale non trattabile in forma chiusa) non sono ravvisabili vantaggi particolari (Baker & Kim 2004).*

[2] *L'uso dei termini difficoltà e abilità è mutuato dall'originario ambito di applicazione dei modelli di Rasch: la valutazione dell'abilità di un gruppo di studenti a cui è sottoposto un test costituito da problemi di difficoltà variabile. Il termine abilità viene comunemente usato anche in altri contesti per indicare la variabile latente d'interesse (per es., nel caso trattato in questa sede, la soddisfazione).*

seguente teorema (si veda Fischer (1995), anche per derivazioni alternative del modello di Rasch):

Teorema 1 *Si supponga di disporre della matrice delle risposte fornite a J item dicotomici da parte di I individui e siano date le seguenti ipotesi:*

1. ***Uni-dimensionalità*** *- I J item sono indicatori della medesima variabile latente θ.*

2. ***Monotonia delle ICC*** *- Le curve caratteristiche di ciascun item $g_j(\theta)$ (Item Characteristic Curve - ICC), cioè le funzioni che esprimono la probabilità di risposta (uguale a 0 o ad 1) al j-esimo item in funzione dell'abilità latente, sono funzioni continue e monotone in senso stretto, decrescenti per la modalità di risposta pari a 0 e crescenti per la modalità pari ad 1.*

3. ***Assenza di guessing*** *- $\lim_{\theta\to-\infty} g_j(\theta) = 0$ e $\lim_{\theta\to\infty} g_j(\theta) = 1$, cioè quanto più il livello di abilità tende a valori piccoli quanto più la probabilità di rispondere "correttamente" al j-esimo item tende a 0, viceversa per livelli di abilità elevati*[3]*.*

4. ***Indipendenza locale degli item*** *- Dato il livello di abilità θ_i per l'individuo i-esimo, le risposte $X_{ij} = x_{ij}$ (con $x_{ij} = 0, 1$) agli item sono tra loro indipendenti:*

$$P[(X_{i1} = x_{i1}) \cap \ldots \cap (X_{iJ} = x_{iJ})|\theta_i] = \Pi_{j=1}^{J} g_j(\theta_i)^{x_{ij}}[1 - g_j(\theta_i)]^{1-x_{ij}}$$

5. ***Sufficienza dei punteggi grezzi*** *- Dato un test di lunghezza J, la statistica dei punteggi grezzi $R_i = \sum_{j=1}^{J} X_{ij}$ è una statistica sufficiente per θ_i.*

---

[3] *La denominazione "assenza di guessing" fa riferimento ai modelli IRT a 3 parametri che prevedono l'introduzione di un asintoto orizzontale per valori di θ tendenti a $-\infty$ che stanno ad indicare la possibilità di ottenere sempre risposte positive agli item come conseguenza del caso.*

*Allora è possibile dimostrare che le Item Characteristic Curves assumono la seguente forma (**Modello di Rasch Dicotomico**):*

$$g_j(\theta_i) = P(X_{ij} = x_{ij}|\theta_i, \beta_j) = \frac{exp[x_{ij}(\theta_i - \beta_j)]}{1 + exp(\theta_i - \beta_j)} \quad (1)$$

Le ipotesi su cui si basa il teorema garantiscono che le stime dei parametri coinvolti (di abilità per le persone e di difficoltà per gli item) abbiano le caratteristiche di uni-dimensionalità e specifica oggettività richieste dal concetto di misura. Si noti, in particolare, che il modello di Rasch è l'unico tra i modelli della famiglia IRT che gode della proprietà di sufficienza dei punteggi grezzi (delle persone e degli item). Ciò significa che, noto il punteggio complessivo che ciascun soggetto ha conseguito nel questionario e che, per definizione, non dipende dalla difficoltà degli item, nessun'altra informazione sull'abilità degli individui è contenuta nei vettori delle risposte: quindi, si verifica facilmente (Wright & Masters (1982) e Baker & Kim (2004)) che la probabilità condizionata del vettore risposta al punteggio complessivo di ciascun individuo dipende soltanto dai parametri di difficoltà degli item e non dai parametri di abilità. E' vero anche il viceversa, cioè la probabilità condizionata al punteggio complessivo di ciascun item dipende soltanto dai parametri di abilità e non dai parametri di difficoltà. Questa importante proprietà va sotto il nome di **separabilità dei parametri** ed è condizione necessaria e sufficiente affinché la specifica oggettività della misura sia garantita (Gori et al. 2005): in concreto, essa afferma infatti che la stima della difficoltà degli item non dipende dall'abilità degli individui che hanno risposto al test e, viceversa, la stima dell'abilità dei soggetti non dipende dallo specifico strumento di misura impiegato.

Al fine di un utilizzo concreto, il modello di Rasch dà origine alle stime, espresse in logit, della difficoltà di ciascun item e dell'abilità di ciascun individuo (con relativi errori standard). Ciò consente di creare una graduatoria di difficoltà degli item e una graduatoria di difficoltà degli individui confrontabili tra loro e al loro interno. In altri termini, è possibile stabilire

quanto un item è più o meno difficile rispetto ad un altro e quanto un soggetto è più o meno abile rispetto ad un altro; è altresì possibile confrontare l'abilità di un individuo con la difficoltà di un item in modo da prevedere la probabilità di scegliere una certa modalità di risposta (ad es., se la difficoltà di un certo item è pari a 0,70 logit, gli individui che presentano un'abilità pari a 0,70 logit hanno una probabilità del 50,0% di scegliere la modalità di risposta 1 piuttosto che 0, mentre per gli individui con abilità pari a 1,3 logit la stessa probabilità sale al 64,5%, per scendere al 35,4% per i soggetti con abilità uguale a 0,1).

Il modello di Rasch si distingue dagli altri modelli IRT (a 2 e a 3 parametri) per il fatto che prevede un solo parametro degli item; per contro, il modello a 2 parametri (2PLM) introduce, accanto al parametro di difficoltà, anche un parametro di discriminazione, che accoglie la possibilità che item diversi abbiano una capacità discriminatoria diversa rispetto alla variabile latente. Concretamente, questo significa che, mentre nel modello di Rasch le ICC sono tra loro parallele e quindi la graduatoria di difficoltà degli item non varia al variare del livello di abilità dei soggetti, nel modello 2PLM le ICC degli item hanno coefficienti di inclinazione differenti e quindi è ammessa la possibilità di graduatorie di difficoltà variabili in funzione del livello di abilità.

L'introduzione di parametri aggiuntivi rende il modello di misura più flessibile e più facilmente adattabile ai dati osservati, ma, per contro, fa venire meno il rispetto delle ipotesi base del modello di Rasch, prima tra tutte la sufficienza dei punteggi complessivi, condizione senza la quale si perde il requisito della specifica oggettività.

Fino ad adesso si è fatto riferimento a test con item dicotomici: l'estensione al caso di item politomici non prevede nessuna modifica del modello dal punto di vista concettuale, ma soltanto l'introduzione di una maggiore complessità dovuta alla presenza di più di due modalità di risposta. La formulazione più generale è data dal Partial Credit Model (Wright &

Masters 1982), di cui il modello di Rasch dicotomico rappresenta un caso particolare. Dal momento che ogni item presenta più di una modalità di risposta (non è richiesto che gli item presentino lo stesso numero di categorie), è necessario stimare un parametro di difficoltà β_{hj} per ogni soglia h di ogni item j, intendendo per soglia il passaggio tra una categoria di risposta e la successiva (dunque il modello dicotomico è un Partial Credit Model dove ogni item presenta una sola soglia a fronte di due modalità di risposta). In generale, le categorie di uno stesso item non saranno ugualmente distanziate, cioè le differenze tra soglie consecutive non saranno costanti, indicando così che la difficoltà di passare da una categoria di risposta alla successiva non è sempre la stessa: per es., date le modalità "decisamente no", "più no che sì", "più sì che no" e "decisamente sì" in un item che misura la soddisfazione per un certo servizio, può darsi che il passaggio da "più no che sì" a "più sì che no" sia più difficile - e quindi la distanza tra le due modalità è maggiore - del passaggio da "più sì che no" a "decisamente sì", fatto questo che indica che quando un individuo è soddisfatto in qualche misura del servizio è anche probabile che ne sia molto soddisfatto. E' invece auspicabile che le soglie siano ordinate, cioè la difficoltà di ogni soglia deve essere maggiore della difficoltà di tutte le soglie precedenti: con riferimento allo stesso esempio, ciò significa che affinché un soggetto scelga la modalità "decisamente sì" e, quindi, superi la terza soglia dell'item, deve anche aver superato le prime due soglie, cioè deve aver preferito la risposta "più no che sì" rispetto a "decisamente no" (superamento della prima soglia) e, poi, la risposta "più sì che no" rispetto a "più no che sì" (superamento della seconda soglia). Il mancato verificarsi di una tale situazione è sintomo di una ridondanza nelle modalità di risposta e viene normalmente risolto procedendo all'aggregazione delle categorie adiacenti (Bond & Fox 2001).

La formula del Partial Credit Model risulta dalla generalizzazione del

modello dicotomico dell'equazione 1:

$$P_{ijx} = P(X_{ij} = x_{ij}|\theta_i, \beta_{jk}) = \frac{exp[\sum_{k=0}^{x_{ij}}(\theta_i - \beta_{jk})]}{\sum_{h=0}^{H_j} exp \sum_{k=0}^{h}(\theta_i - \beta_{jk})} \qquad (2)$$

con $x_{ij} = 0, 1, \ldots, h, \ldots, H_j$.

2.3 LA STIMA DEI PARAMETRI

In letteratura sono noti tre principali metodi di stima dei parametri del modello di Rasch (Wright & Masters (1982), Molenaar (1995) e Baker & Kim (2004)): la massima verosimiglianza congiunta o non condizionata (*Joint Maximum Likelihood* - JML), la massima verosimiglianza condizionata (*Conditional Maximum Likelihood* - CML) e la massima verosimiglianza marginale (*Marginal Maximum Likelihood* - MML).

Il metodo della massima verosimiglianza congiunta procede alla stima simultanea dei parametri di abilità e difficoltà, attraverso la massimizzazione della funzione di log-verosimiglianza, che nel caso dicotomico assume la seguente forma:

$$L = \sum_{i=1}^{I} r_i\theta_i - \sum_{j=1}^{J} s_j\beta_j - \sum_{i=1}^{I}\sum_{j=1}^{J} \log[1 + exp(\theta_i - \beta_j)] \qquad (3)$$

dove: $r_i = \sum_{j=1}^{J} x_{ij}$ e $s_j = \sum_{i=1}^{I} x_{ij}$.

Siccome non è possibile pervenire ad una soluzione finita delle equazioni di stima (ottenute dall'imposizione delle condizioni del primo ordine), è necessario ricorrere ad una procedura iterativa tramite algoritmo di Newton-Raphson; l'equazione risolutiva che si ottiene alla t-esima iterazione è data da:

$$\begin{bmatrix} \widehat{\theta}_i \\ \widehat{\beta}_j \end{bmatrix}_{t+1} = \begin{bmatrix} \widehat{\theta}_i \\ \widehat{\beta}_j \end{bmatrix}_t - \begin{bmatrix} \widehat{L}_{ii} & \widehat{L}_{ij} \\ \widehat{L}_{ij} & \widehat{L}_{jj} \end{bmatrix}_t^{-1} \times \begin{bmatrix} \widehat{L}_i \\ \widehat{L}_j \end{bmatrix}_t$$

dove con $\widehat{L}_{..}$ sono state indicate le derivate seconde e con $\widehat{L}_{.}$ le derivate prime.

Il principale difetto della procedura JML è dovuto al fatto che le stime che si ottengono non sono consistenti per questionari con un numero J di item finito: la consistenza si ha soltanto per $I \longrightarrow \infty$, $J \longrightarrow \infty$ e $I/J \longrightarrow \infty$ (Molenaar (1995) e Baker & Kim (2004)).

Gli altri due metodi di stima, invece, forniscono stime consistenti anche per $I \longrightarrow \infty$ con ampiezza J del questionario finita. In particolare, il metodo CML procede alla massimizzazione della funzione di log-verosimiglianza *condizionata* al punteggio complessivo di ciascun individuo $r_i = \sum_{j=1}^{J} x_{ij}$:

$$L = -\sum_{j=1}^{J} s_j \beta_j - \sum_{r=1}^{I-1} f_r \log \gamma(r, \beta) \tag{4}$$

dove: f_r indica il numero di individui che hanno conseguito un punteggio pari ad r e $\gamma(r, \beta)$ è la cosiddetta funzione simmetrica pari a

$$\sum_{x_{ij}=1}^{r} exp(-\sum_{j=1}^{J} x_{ij}\beta_j).$$

In virtù della sufficienza dei punteggi r_i la funzione da massimizzare dipende soltanto dai parametri di difficoltà degli item, che quindi sono gli unici elementi presenti nelle equazioni di stima, anch'esse risolvibili iterativamente tramite procedura di Newton-Raphson:

$$\begin{bmatrix} \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_{J-1} \end{bmatrix}_{t+1} = \begin{bmatrix} \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_{J-1} \end{bmatrix}_{t} - \begin{bmatrix} \widehat{L}_{11} & \dots & \widehat{L}_{1(J-1)} \\ \vdots & \vdots & \vdots \\ \widehat{L}_{(J-1)1} & \dots & \widehat{L}_{(J-1)(J-1)} \end{bmatrix}_{t}^{-1} \times \begin{bmatrix} \widehat{L}_1 \\ \vdots \\ \widehat{L}_{J-1} \end{bmatrix}$$

Una volta ottenuti i parametri degli item, i relativi valori vengono impiegati per stimare i parametri di abilità, che quindi richiedono una procedura separata.

Infine, sotto l'approccio della massima verosimiglianza marginale, si assume che i soggetti rappresentino un campione casuale da una popolazione la cui abilità è distribuita secondo una certa funzione di densità, $g(\theta|\tau)$, con τ vettore dei parametri di $g(.)$. In questa situazione, dunque, i parametri degli item sono considerati effetti fissi, mentre le abilità sono effetti

casuali: integrando la funzione di verosimiglianza sulla distribuzione di abilità, i parametri casuali di abilità vengono rimossi e così i parametri degli item possono essere stimati in maniera consistente per qualunque ampiezza campionaria, dal momento che l'incremento del numero di individui non determina un aumento nel numero di parametri della popolazione. In concreto, applicando il teorema di Bayes, si ottiene la seguente relazione:

$$P(\theta_i|x_{i.},\beta,\tau) = \frac{P(x_{i.}|\theta_i,\beta)g(\theta_i|\tau)}{\int_\Theta P(x_{i.}|\theta_i,\beta)g(\theta_i|\tau)d\theta_i} \tag{5}$$

Il denominatore dell'equazione 5 è la probabilità marginale del vettore risposta agli item $x_{i.}$ relativo all'i-esimo individuo rispetto ai parametri degli item e alla densità di abilità della popolazione. Sommando rispetto all'insieme degli I individui si ottiene la funzione di log-verosimiglianza marginale da massimizzare:

$$L = \sum_{i=1}^{I} \log \int_\Theta P(x_{i.}|\theta_i,\beta)g(\theta_i|\tau)d\theta_i$$

Derivando rispetto al vettore β dei parametri di difficoltà e al vettore τ dei parametri della distribuzione di abilità[4], si ottengono le equazioni di verosimiglianza la cui soluzione fornisce le stime desiderate. Nell'approccio di Bock e Lieberman (Baker & Kim 2004), le equazioni di stima vengono risolte ricorrendo all'approssimazione integrale di Gauss-Hermite e il metodo dello *scoring* di Fisher viene impiegato per stimare simultaneamente i parametri. Lo svantaggio principale di tale approccio è la pesantezza computazionale: la stima simultanea dei J parametri di difficoltà richiede l'inversione di una matrice d'informazione di dimensione $J \times J$, rivelandosi quindi fattibile soltanto per questionari di dimensione limitata. Per

[4] *Affinché il modello sia identificabile è necessario porre dei vincoli: solitamente o si stabilisce che la somma dei parametri di difficoltà è pari a 0 e, di conseguenza, verranno stimati (J-1) elementi di β e tutti gli elementi di τ, oppure un elemento di τ viene vincolato a 0 (se g(.) è una distribuzione normale si pone la media pari a 0) e, quindi, vengono stimati tutti gli elementi del vettore β e gli elementi residui di τ.*

contro, l'approccio sviluppato successivamente da Bock e Aitkin (Baker & Kim 2004) e basato sull'algoritmo EM per la soluzione delle equazioni di verosimiglianza è computazionalmente molto più efficiente, in quanto consente di stimare i parametri di difficoltà un item alla volta: infatti, poiché viene assunta l'indipendenza degli item (oltre che l'indipendenza degli individui e l'indipendenza degli item e degli individui), le derivate seconde miste sono pari a zero e, dunque, la fase di massimizzazione (step M) del valore atteso della funzione di densità congiunta a posteriori viene effettuata per ogni item singolarmente.

Una volta ottenute le stime di difficoltà, la stima dell'abilità degli individui può essere ricavata in base a varie procedure.

Lo stimatore atteso a posteriori (*Expected a posteriori* - EAP) è dato dal valore atteso della funzione di densità a posteriori di θ, condizionata ai vettori $x_{i.} = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ e $\beta = (\beta_1, \beta_2, \ldots, \beta_J)$:

$$E(\theta_i|x_{i.}, \widehat{\beta}, \widehat{\tau}) = \int_{\Theta} \theta_i P(\theta_i|x_{i.}, \widehat{\beta}, \widehat{\tau}) d\theta = \frac{\int_{\Theta} \theta_i P(x_{i.}|\theta, \widehat{\beta}) g(\theta|\widehat{\tau}) d\theta}{\int_{\Theta} P(x_{i.}|\theta, \widehat{\beta}) g(\theta|\widehat{\tau}) d\theta}$$

Lo stimatore modale di Bayes è invece ottenuto massimizzando la funzione di densità a posteriori di θ, condizionata su $x_{i.}$ e β, rispetto a θ:

$$\max \left[P(\theta|x_{i.}, \widehat{\beta}, \widehat{\tau}) \right] = \max \left[\frac{\theta P(x_{i.}|\theta, \widehat{\beta}) g(\theta|\widehat{\tau})}{\int_{\Theta} P(x_{i.}|\theta, \widehat{\beta}) g(\theta|\widehat{\tau}) d\theta} \right]$$

Infine, lo stimatore di massima verosimiglianza deriva dalla massimizzazione della funzione di log-verosimiglianza $L = \log P(x_{i.}|\theta, \widehat{\beta})$.

2.4 STATISTICHE DI ADATTAMENTO E DIF

Una volta ottenute le stime dei parametri, il confronto tra risposte osservate e valori attesi consente di esprimere un giudizio sulla bontà di adattamento del modello impiegato ai dati osservati. Nell'ambito dei modelli di Rasch è particolarmente utile valutare la bontà di adattamento, in modo da individuare eventuali violazioni delle ipotesi fondanti il modello.

Le statistiche maggiormente impiegate per la diagnostica del modello di Rasch sono le statistiche **Outfit** e **Infit**, basate sul confronto tra risposte osservate per ciascun individuo a ciascun item del questionario e risposte attese sulla base del modello di Rasch stimato (Wright & Masters 1982).

Indicando con P_{ijx} la probabilità che l'individuo i-esimo scelga la categoria x per il j-esimo item e con X_{ij} la risposta osservata per l'individuo i-esimo e l'item j-esimo, si ha che:

$$E_{ij} = \sum_{x=0}^{r_j} x P_{ijx}$$

è il valore atteso della risposta;

$$Y_{ij} = X_{ij} - E_{ij}$$

è il residuo corrispondente;

$$W_{ij} = \sum_{x=0}^{r_j} (X_{ij} - E_{ij})^2 P_{ijh}$$

è la varianza della risposta osservata X_{ij} ed assume valore massimo quando le stime di abilità e difficoltà sono identiche, mentre tende a ridursi all'aumentare della differenza in valore assoluto tra difficoltà dell'item j e abilità dell'individuo i;

$$Z_{ij} = Y_{ij}/\sqrt{W_{ij}}$$

è il residuo standardizzato avente una distribuzione Normale con media pari a 0 e varianza unitaria.

Per valutare l'adattamento complessivo di un item al modello di Rasch si calcola la media aritmetica semplice o ponderata dei residui standardizzati al quadrato. In particolare, nel caso di una media aritmetica semplice si ottiene la **statistica Outfit** (o ***Unweighted Mean Square statistic***):

$$OUT_j = \sum_{i=1}^{I} Z_{ij}^2 / I,$$

avente varianza pari a:

$$s_j^2 = \sum_{i=1}^{I} \left((\sum_{x=0}^{r_j} (x - E_{ij})^4 P_{ijh}) / W_{ij}^2 \right) / I^2 - 1/I$$

Dal momento che s_j varia in funzione del numero di individui nel campione e W_{ij} varia sia da item ad item che da campione a campione, non è facile determinare un livello di *cut-off* generale per valutare la bontà di adattamento di un item; di conseguenza, solitamente si procede alla standardizzazione (trasformazione di Wilson-Hilferty), ottenendo una statistica con distribuzione approssimativamente Normale con media 0 e varianza unitaria:

$$t_j OUT = (OUT_j^{1/3} - 1)(3/s_j) + (s_j/3)$$

Poiché la statistica Outfit è il risultato di una media aritmetica semplice, essa risulta particolarmente sensibile a risposte inattese (cioè improbabili) che provengono da individui per i quali l'item j risulta inappropriato, in quanto troppo facile o troppo difficile. Per ovviare a questo problema la statistica Outfit viene solitamente affiancata (o sostituita) con la **statistica Infit** (o ***Weighted Mean Square statistic***) che pondera i residui standardizzati al quadrato con le rispettive varianze individuali:

$$IN_j = \sum_{i=1}^{I} W_{ij} Z_{ij}^2 / \sum_{i=1}^{I} W_{ij}$$

con varianza:

$$q_j^2 = \sum_{i=1}^{I} \left((\sum_{x=0}^{r_j} (h - E_{ij})^4 P_{ijx}) / W_{ij}^2 \right) / \left(\sum_{i=1}^{I} W_{ij}^2 \right)$$

Dal momento che la varianza W_{ij} è tanto maggiore quanto più le stime di abilità e di difficoltà sono simili, la statistica Infit dà maggiore peso alle risposte degli individui per i quali l'item j è ben calibrato, cioè ha un livello di difficoltà in linea con l'abilità del soggetto. Anche in tal caso, in pratica, si utilizza la statistica Infit standardizzata:

$$t_j IN = (IN_j^{1/3} - 1)(3/q_j) + (q_j/3),$$

che ad un livello di significatività del 5% assume valori compresi nell'intervallo $[-2; +2]$.

La prassi consiste nell'eliminare in una procedura iterativa tutti gli item (e gli individui[5]) che presentano un cattivo adattamento al modello, cioè valori che fuoriescono dal suddetto intervallo di significatività. Spesso l'insieme di item esclusi contribuisce a misurare una dimensione separata; nei casi più estremi, invece, può accadere che non sia possibile individuare nessun insieme di item coerenti con le ipotesi del modello di Rasch: questo può essere causato o da un questionario mal calibrato oppure da un miscuglio di individui apparentemente appartenenti alla stessa popolazione, ma in realtà afferenti a popolazioni diverse. Quest'ultimo caso può essere sintomo di un funzionamento diverso degli item in corrispondenza di gruppi di individui distinti: tale fenomeno va sotto il nome di ***Differential Item Functioning*** o **DIF**.

Più precisamente, un item è considerato distorto se, condizionatamente ad un certo livello di abilità, la probabilità di risposta corretta (nel caso di item dicotomico) o, più in generale, la probabilità di scegliere una certa modalità di risposta differisce in maniera sistematica tra sottogruppi di individui (per es., tra maschi e femmine, tra studenti di corsi di laurea diversi, ecc.). La presenza di uno o più item distorti in un questionario fa venire meno il rispetto del principio di specifica oggettività. Infatti, se per due individui a e b la difficoltà di uno stesso item è diversa, β_{aj} e β_{bj}, può accadere che il soggetto con abilità minore superi l'item con maggiore probabilità del soggetto con abilità maggiore e il confronto di abilità tra i due non risulta più indipendente dallo specifico item impiegato (Gori et al. 2005):

$$\log \frac{P(X_{aj}=1)}{P(X_{aj}=0)} - \log \frac{P(X_{bj}=1)}{P(X_{bj}=0)} = (\theta_a - \beta_{aj}) - (\theta_b - \beta_{bj}) \neq (\theta_a - \theta_b)$$

[5] *Le statistiche Infit ed Outfit possono essere calcolate, con procedura analoga, anche per gli individui.*

L'impatto del DIF sulla validità di un questionario e, conseguentemente, sull'oggettività delle misure dipende sia dal numero di item distorti sia dall'entità delle differenze nei parametri di difficoltà per i vari item tra i diversi sottogruppi della popolazione. In letteratura esistono varie proposte per la diagnostica del DIF (Glas & Verhelst 1995), ma quella più diffusa e implementata nei software di uso più comune (Wu, Adams & Wilson (1998) e Tesio, Valsecchi, Sala, Guzzon & Battaglia (2002)) si basa sull'analisi dei residui tra i sottogruppi individuati rispetto a una o più variabili di aggregazione. In particolare, il software ConQuest, impiegato per l'analisi presentata nel paragrafo successivo, tramite l'inserimento nel modello di un'interazione tra ciascun item (o ciascuna modalità di risposta per ogni item, nel caso di modelli per item politomici) e la variabile di aggregazione, perviene alla stima degli effetti differenziali positivi o negativi rispetto alla difficoltà media dell'item e tramite il classico test χ^2 valuta la significatività statistica di tali differenze.

Nel caso in cui la presenza di DIF sia significativa, il ricorso a modelli IRT a più parametri può rappresentare una soluzione: infatti, nel modello a due parametri l'inserimento di un parametro di discriminazione consente di prevedere in modo esplicito la possibilità che uno stesso item discrimini in maniera diversa per livelli di abilità differenti. In alternativa, volendo rimanere nel contesto del modello di Rasch (unica soluzione ammissibile se ci si vuole attenere al concetto di misura quale definito in questa sede), sarà necessario individuare gruppi omogenei di individui rispetto ai quali il modello di Rasch presenta un buon adattamento.

3. VALUTAZIONE DEL SISTEMA UNIVERSITARIO ATTRAVERSO LE OPINIONI DEGLI STUDENTI

3.1 GLI STUDENTI DELL'ATENEO FIORENTINO E LA VALUTAZIONE DELLA DIDATTICA

Dopo aver sinteticamente richiamato nel paragrafo precedente le assunzioni e la struttura dei modelli di Rasch, in questo paragrafo l'interesse si concentra sull'utilità di questa tipologia di modelli al fine della valutazione delle performance di un sistema complesso, quale quello universitario.

I dati utilizzati nell'analisi svolta sono relativi agli studenti frequentanti dell'Università di Firenze che, negli anni 2003, 2004 e 2005, hanno compilato il questionario sulla valutazione della didattica relativo a singoli insegnamenti, il cui testo è riportato in appendice A. I questionari raccolti ammontano ad un totale di 237133[6], suddivisi piuttosto equamente tra i 3 anni di rilevazione (71262 per il 2003, 72509 per il 2004 e 93362 per il 2005). E' interessante osservare come gli anni presi in esame siano quelli immediatamente successivi alla riforma dei cicli e degli ordinamenti didattici e, dunque, i diversi questionari possono essere ricondotti a tipologie di corsi di laurea tra loro molto differenti: accanto alle lauree (ad esaurimento) del vecchio ordinamento (23051 questionari raccolti), si trovano infatti le lauree di primo livello (192763 questionari), le lauree specialistiche di secondo livello (601 questionari) e le lauree specialistiche a ciclo unico (20718 questionari). Questa eterogeneità dei corsi di laurea riflette una più generale eterogeneità a livello di caratteristiche degli studenti presi in considerazione: come sarà meglio evidenziato dalle analisi successive, ciò si ripercuote sulla bontà di adattamento del modello di Rasch e sulla necessità di tenere esplicitamente in considerazione la struttura complessa della popolazione.

---

[6] *A questi se ne aggiungono altri 36198 che, però, non essendo attribuibili a nessun corso di laurea specifico non sono stati presi in considerazione.*

Il questionario utilizzato è relativo all'opinione degli studenti in termini di soddisfazione per aspetti specifici dell'insegnamento e, più in generale, del corso di laurea frequentati. Oltre ad alcune informazioni generali sulle caratteristiche dello studente intervistato (tipo di maturità conseguita, anno d'iscrizione, frequenza del corso) e alla possibilità di suggerimenti in forma chiusa ed aperta, il corpo centrale del questionario è costituito da 22 domande relative alla soddisfazione per diversi aspetti dell'insegnamento o del corso di studi (organizzazione, docenza, aule, aspetti specifici del corso di studi, altre informazioni) più un'ultima domanda inerente la soddisfazione globale per l'insegnamento. Ciascun item è costituito da 4 modalità di risposta ordinali, di cui due denotano un giudizio negativo e le altre due un giudizio positivo. La struttura del questionario è, dunque, tipica per l'applicazione di un modello di Rasch, dove la variabile latente oggetto di misurazione è la soddisfazione degli studenti per i vari insegnamenti e corsi di laurea.

La condizione necessaria per l'utilizzo dei modelli di Rasch è l'individuazione di gruppi omogenei di studenti, tali per cui abbia senso applicare il concetto di misura. Al fine di un'utilità concreta dei risultati ottenuti dalla stima del modello è poi ulteriormente necessario che tali gruppi omogenei siano riconducibili a centri decisionali ben precisi: nell'ambito universitario si tratterà, ad es., di singoli Atenei, facoltà, corsi di laurea ed insegnamenti. Una volta individuata la tipologia di centro decisionale rispetto a cui interessa effettuare la misura, la stima del modello di Rasch si risolve in due diversi tipi di graduatorie:

- Una graduatoria di abilità o, visto il contesto di riferimento, di soddisfazione, che consente di quantificare per ogni gruppo omogeneo di studenti il livello di soddisfazione e di effettuare confronti rispetto agli altri gruppi considerati. Nel caso in cui si disponga di osservazioni su più anni è altresì possibile valutare eventuali cambiamenti intervenuti nelle posizioni in graduatoria. Tale graduatoria può, dunque, essere

considerata uno strumento di confronto *tra* gruppi.

- Una graduatoria di difficoltà degli item, che permette di individuare per ogni gruppo omogeneo di studenti gli elementi più critici, cioè gli elementi rispetto ai quali è improbabile osservare persone soddisfatte o, con linguaggio più tecnico, il cui superamento (risposta 1 piuttosto che 0, nel caso dicotomico) richiede un livello di soddisfazione elevato. Si tratta, in tal caso, di uno strumento di monitoraggio *interno* ai gruppi.

I due tipi di graduatoria non sono strumenti indipendenti tra loro, al contrario la loro utilità si sostanzia in un utilizzo congiunto dei due. In particolare, la posizione in graduatoria occupata da ciascun centro decisionale e le eventuali variazioni che essa subisce nel corso del tempo, possono essere spiegate, almeno in parte, tramite gli effetti di interventi o di mancati interventi sui punti critici evidenziati dalle graduatorie di difficoltà. Si puntualizza che, nell'analisi dei fattori rispetto ai quali gli studenti sono meno soddisfatti, è d'interesse per il centro decisionale individuare quelli su cui esso ha un effettivo potere d'intervento (ad es. il singolo docente può agire sull'adeguatezza del materiale didattico, ma non sull'adeguatezza delle aule in cui si svolgono le lezioni, perlomeno non in maniera diretta), tenendo conto che, comunque, parte delle variazioni nella posizione in graduatoria nel corso del tempo possono essere imputabili anche a cambiamenti intervenuti in altri gruppi o a cambiamenti nella popolazione di studenti in termini di composizione o aspettative (anche se quest'ultima ipotesi dovrebbe essere piuttosto improbabile dal momento che le rilevazioni hanno cadenza annuale).

Quindi, il modello di Rasch consente non solo di ottenere graduatorie di merito per valutare la performance di un insieme di centri decisionali, ma fornisce altresì uno strumento di supporto all'individuazione di oppor-

tune politiche d'intervento da attuarsi in futuro e di verifica dell'esito di interventi già posti in essere in passato.

3.2 ANALISI EMPIRICA

L'analisi iniziale è stata condotta sul data set del 2005 privo di risposte mancanti, prendendo in esame le domande da d1 a d11 e da d17 a d19; le domande da d12 a d16 sono state escluse in quanto diverse tra le varie facoltà, mentre la d23 relativa alla soddisfazione complessiva è stata considerata separatamente. Per la stima dei modelli di Rasch è stato utilizzato il software ConQuest (Wu et al. 1998), che ricorre al metodo di stima della massima verosimiglianza marginale con applicazione dell'algoritmo sviluppato da Bock e Aitkin.

I risultati relativi alle stime dei parametri di difficoltà degli item riportati in Tab. 1 e derivanti da un modello di Rasch dicotomico[7] mostrano un pessimo adattamento dei dati al modello. Di norma, valori delle statistiche standardizzate Outfit ed Infit fuori dai limiti di significatività possono essere spiegati attraverso la presenza di multidimensionalità. In tali casi è opportuno individuare sottoinsiemi di item unidimensionali rispetto ai quali il modello di Rasch presenta un buon adattamento: questo può essere fatto eliminando in successione gli item con un cattivo adattamento oppure avvalendosi del supporto di altre tecniche, quali l'analisi fattoriale e l'analisi dell'andamento del coefficiente α di Cronbach. In tale caso, però, l'analisi di dimensionalità non ha condotto a risultati soddisfacenti. Un'altra causa del cattivo adattamento del modello può essere ricercata in una eccessiva disomogeneità dell'insieme di studenti analizzato, tale da determinare un effetto DIF consistente. A questo proposito, se si considera l'aggregazione per tipologia di corso di laurea (lauree di primo livello, di secondo livello, specialistiche a ciclo unico e lauree del vecchio ordinamento) si osserva

---

[7] *La stima di un Partial Credit Model ha condotto a risultati analoghi, con l'unica differenza di una maggiore lentezza computazionale.*

(cfr. Tab. 2) che non solo sussiste una differenza significativa nel livello di soddisfazione medio per tipo di corso di laurea, ma c'è evidenza di una significativa distorsione degli item: in altri termini, il questionario funziona in modo diverso a seconda dell'aggregazione di corsi di laurea considerata, dando origine a graduatorie di difficoltà degli aspetti esaminati tra loro differenti.

Tab. 1: Popolazione degli studenti frequentanti, anno 2005: stima di difficoltà degli item, errore standart, statistiche di adattamento standardizzante Outfit e Infit.

Item	Stima	E.S.	Stand. Outfit	Stand. Infit
d1	-0,974	0,014	-1,2	3,5
d2	-0,812	0,014	-2,3	1,1
d3	-1,712	0,015	0,0	0,6
d4	-1,564	0,015	-9,4	-7,6
d5	-2,087	0,016	-8,4	-6,4
d6	-1,653	0,016	-3,1	-0,9
d7	-2,600	0,020	-3,1	-2,8
d8	-2,821	0,021	-22,3	-9,6
d9	-1,916	0,016	-18,4	-15,9
d10	-1,977	0,016	-17,6	-13,4
d11	-2,792	0,021	-29,1	-11,0
d17	-1,026	0,014	5,1	16,1
d18	-0,815	0,014	3,3	12,9
d19	-1,488	0,014	5,7	9,5
d20	-0,978	0,014	10,9	17,8
d21	-1,557	0,015	24,7	28,4
d22	-2,664	0,019	-11,2	-4,5

Tab. 2: Popolazione degli studenti frequentanti, anno 2005: test Chi-quadrato per il confronto della soddisfazione e per la significatività del DIF rispetto al tipo di corso di laurea.

Var.raggruppamento	Chi-quadrato	GdL	Prob.
Tipo di corso di laurea	652,5	3	0,000
Item*Tipo di corso di laurea	2413,1	51	0,000

I risultati ottenuti indicano chiaramente che la via più opportuna da perseguire è quella di un'analisi separata per ciascuna tipologia di corsi di laurea. Si sono pertanto prese in considerazione le quattro lauree specialistiche a ciclo unico (Architettura, Chimica e tecnologie farmaceutiche, Farmacia, Medicina e chirurgia), il modello di Rasch stimato[8] presenta un adattamento decisamente migliore (cfr. Tab. 3), anche se ancora numerosi item mostrano valori delle statistiche Outfit ed Infit non significativi al livello del 5%. Anche in tal caso l'analisi di dimensionalità non consente di pervenire a risultati migliori: come verificato tramite il test Chi-quadrato (cfr. Tab. 4), sussiste ancora un DIF significativo rispetto alla variabile corso di laurea. Scendendo in maggiore dettaglio, le stime di soddisfazione media dei 4 corsi di laurea esaminati e presentate in Tab. 5 risultano significativamente diverse: in particolare, i meno soddisfatti sono gli studenti di Medicina e chirurgia, mentre i più soddisfatti sono i colleghi di Farmacia. Tali risultati sono, inoltre, coerenti con quanto deriva dall'esame della domanda di soddisfazione globale (ultima colonna di Tab. 5).

A questo punto è utile cercare di capire a quali conclusioni errate si può pervenire ignorando la presenza di DIF e il conseguente cattivo adattamento del modello al data set. In Tab. 6 sono riportate le graduatorie di difficoltà

[8] *A causa dell'elevato numero di mancate risposte alle domande d5 e d18, da attribuire al fatto che le attività didattiche integrative, a cui tali quesiti fanno riferimento, in molti insegnamenti non sono previste, tali domande sono state escluse dall'analisi.*

Tab. 3: Popolazione degli studenti frequentanti di Lauree Specialistiche a Ciclo Unico (Lscu), anno 2005: stima di difficoltà degli item, errore standart, statistiche di adattamento standardizzante Outfit e Infit.

Item	Stima	E.S.	Stand. Outfit	Stand. Infit
d1	-0,724	0,028	0,0	1,5
d2	-0,863	0,028	-0,8	-1,0
d3	-1,443	0,031	1,4	1,2
d4	-1,292	0,030	-3,5	-4,0
d6	-1,292	0,030	0,2	-0,2
d7	-2,868	0,045	-2,7	-0,7
d8	-2,627	0,043	-7,2	-2,2
d9	-1,409	0,032	-5,2	-5,4
d10	-1,687	0,034	-6,0	-5,1
d11	-2,595	0,043	-8,9	-2,6
d17	-1,372	0,029	4,5	8,7
d19	-1,053	0,029	2,0	4,0
d20	-0,539	0,028	3,0	7,6
d21	-0,994	0,029	6,4	11,2
d22	-2,235	0,038	-3,2	-1,2

Tab. 4: Popolazione degli studenti frequentanti di Lauree Specialistiche a Ciclo Unico (Lscu), anno 2005: test Chi-quadrato per il confronto della soddisfazione e per la significatività del DIF rispetto al corso di laurea.

Var.raggruppamento	Chi-quadro	GdL	Prob.
Corso di laurea	382,6	3	0,000
Item*Corso di laurea	3345,2	45	0,000

Tab. 5: Popolazione degli studenti frequentanti di Lauree Specialistiche a Ciclo Unico (Lscu), anno 2005: stima di errore standart della soddisfazione media per corso di laurea, percentuale di soddisfatti (più si che no e decisamente si) in base alla domanda d23.

Corso di laurea	Stima	E.S.	% d23
Farmacia	0,176	0,013	87,7
Architettura	0,114	0,009	87,1
Chimica e tecn.farm.	0,081	0,013	77,7
Medicina e chirurgia	-0,371	–	72,4

degli item per ciascuno dei 4 corsi di laurea esaminati separatamente e la graduatoria "media" risultante dal modello stimato sull'intero data set (cfr. Tab. 3).

Innanzitutto, per alcuni item è possibile osservare una unanimità di giudizio tra i 4 corsi di laurea: gli aspetti rispetto ai quali gli studenti si ritengono soddisfatti con maggiore probabilità sono quelli relativi al rispetto degli orari delle lezioni (d7), alla reperibilità del docente per chiarimenti e spiegazioni (d8) e alla disponibilità del docente a spiegazioni esaurienti (d11); per contro, gli elementi più critici riguardano l'adeguatezza del carico di lavoro complessivo degli insegnamenti previsti nel medesimo periodo di riferimento (d1) e la regolare attività di studio durante la frequenza delle lezioni (d20). Quest'ultimo aspetto, benché non direttamente controllabile dal docente, può comunque essere parzialmente influenzato agendo su altri elementi critici, quali il carico di studio complessivo (d1) oppure fornendo conoscenze preliminari più adeguate all'impegno richiesto dall'insegnamento (d19) o ancora innovando il contenuto dell'insegnamento (d21). Probabilmente, invece, stimolare l'interesse degli studenti verso la disciplina (d9 e d22) non è sufficiente per indurre gli studenti ad un'attività di studio più regolare, dal momento che entrambi gli item non presentano un livello di criticità particolare.

Al di là di queste considerazioni senz'altro rilevanti che scaturiscono dall'analisi della graduatoria complessiva dei corsi di laurea specialistica a ciclo

unico, l'analisi delle graduatorie di difficoltà dei singoli corsi di laurea pone in evidenza alcune peculiarità che spingono ad ulteriori riflessioni su aspetti critici specifici di un corso di studi e non di altri. A titolo esemplificativo, l'adeguatezza delle aule (d17) a livello complessivo occupa una posizione intermedia nella graduatoria, denotando così una situazione che, pur presentando margini di miglioramento per incrementare la soddisfazione degli studenti, non si configura come elemento prioritario. Approfondendo l'analisi per i singoli corsi di laurea, si osserva invece che, mentre per gli studenti di Chimica e tecnologie farmaceutiche il giudizio sulle aule è decisamente più positivo rispetto alla media, per gli studenti di Architettura diventa l'aspetto di maggiore insoddisfazione su cui, dunque, sarebbe opportuno concentrare gli sforzi di miglioramento. Considerazioni analoghe possono essere svolte per altri item, quali l'adeguatezza delle conoscenze preliminari per la comprensione degli argomenti trattati nei singoli insegnamenti (d19): se da una parte per gli studenti di Architettura, Chimica e tecnologie farmaceutiche e Farmacia questo elemento rappresenta uno dei principali punti critici, dall'altra il giudizio degli studenti di Medicina e chirurgia è decisamente migliore. Al contrario, l'opinione di questi ultimi sulla chiarezza delle modalità di esame è peggiore di quella dei colleghi di Architettura e Chimica e tecnologie farmaceutiche.

Un'ultima considerazione riguarda il campo di variazione delle stime di difficoltà per ciascun corso di laurea: questo è massimo e pari a 3.233 logit per Chimica e tecnologie farmaceutiche e minimo e pari a 1.633 logit per Medicina e chirurgia, mentre per Architettura e Farmacia assume un valore intermedio e uguale rispettivamente a 2.537 e 2.778 logit. Si può, quindi, affermare che per Medicina e chirurgia la distanza in termini di difficoltà tra item è inferiore a quella che si osserva negli altri corsi, mentre a Chimica e tecnologie farmaceutiche il questionario riesce a coprire un intervallo di soddisfazione più ampio rispetto agli altri corsi. In generale, benché non sia possibile definire dei valori ottimali a priori, per un buon funzionamento

del questionario è auspicabile che la difficoltà minima e massima stimate siano tali da comprendere i livelli di soddisfazione minimi e massimi osservabili nella popolazione e, all'interno di questo intervallo, si richiede che la distanza tra un parametro di difficoltà e il successivo non sia né troppo ampia (altrimenti non si riuscirebbero a misurare livelli di soddisfazione intermedi) né troppo ridotta (nel caso estremo di due item con uguale stima di difficoltà è evidente che uno dei due è ridondante). Sulla questione dell'adeguatezza del questionario in funzione della distribuzione di soddisfazione della popolazione si torna al termine del paragrafo.

Al di là delle considerazioni svolte fino ad ora, il problema iniziale del cattivo adattamento del modello di Rasch ai dati non è ancora stato risolto, in quanto non si è pervenuti all'individuazione di una popolazione sufficientemente omogenea. Ricapitolando, è stato posto in evidenza un funzionamento distorto degli item sia a livello di tipologie di corsi di laurea sia, relativamente alle lauree specialistiche a ciclo unico, a livello di corsi di laurea. E', dunque, opportuno spostare l'analisi ad un livello di aggregazione più basso, quello del singolo insegnamento: a titolo esemplificativo verrà considerato il corso di studi in Architettura. Prendendo in esame i 14 insegnamenti attivati negli anni 2003, 2004 e 2005 ad Architettura[9] e

---

[9] *La codifica ufficiale adottata per gli insegnamenti è la seguente:*
- *29101: Tecnologia dei materiali*
- *20509: Laboratorio di tecnologia*
- *29341: Tecnologia dell'architettura*
- *29173: Tecnologia dell'architettura*
- *29111: Tecnologie per le energie rinnovabili*
- *25005: Analisi del territorio e insediamenti*
- *29083: Restauro archeologico*
- *25001: Disegno dell'architettura*
- *20593: Laboratorio di restauro*
- *29117: Laboratorio di costruzioni II*
- *29236: Storia e metodi dell'architettura*
- *29250: Fisica tecnica ambientale*
- *29181: Laboratorio di costruzioni II*
- *29009: Analisi del territorio e insediamenti.*

Si noti che alcuni insegnamenti hanno la medesima denominazione, ma codifiche diverse, in quanto tenuti da docenti differenti: è, dunque, corretto considerarli come insegnamenti a sé.

aggiungendo le domande d12 “adeguatezza del comportamento del docente nei riguardi degli studenti” e d14 “trattamento esauriente degli argomenti affrontati alle lezioni” (in quanto comuni a tutta la facoltà e non influenzati da troppe risposte mancanti), i risultati del test Chi-quadrato in Tab. 7 mostrano che il DIF è presente anche a livello di insegnamento, per ciascuno dei tre anni considerati.

Tab. 6: Popolazione degli studenti frequentanti di Lauree Specialistiche a Ciclo Unico (Lscu), anno 2005: graduatoria di difficoltà degli item, per corso di laurea e complessiva.

Architettura		Chimica e Tecn.farm.		Farmacia		Medicina e chirurgia		Class.globale	
Item	Stima	Item	Stima	Item	Stima	Item	Stima	Item	Stima
d17	-0,178	d1	-0,166	d20	-0,284	d20	-0,850	d20	-0,539
d1	-0,679	d2	-0,257	d21	-0,514	d1	-0,898	d1	-0,724
d19	-0,681	d20	-0,292	d19	-1,057	d6	-1,032	d2	-0,863
d20	-0,730	d19	-0,844	d1	-1,153	d2	-1,071	d21	-0,994
d2	-0,734	d9	-1,002	d6	-1,175	d4	-1,175	d19	-1,053
d21	-0,990	d21	-1,015	d4	-1,329	d3	-1,186	d4	-1,292
d4	-1,193	d10	-1,240	d2	-1,390	d9	-1,293	d6	-1,292
d6	-1,356	d4	-1,471	d3	-1,494	d17	-1,308	d17	-1,372
d3	-1,564	d3	-1,528	d9	-1,527	d21	-1,457	d9	-1,409
d9	-1,814	d6	-1,605	d17	-1,666	d19	-1,630	d3	-1,443
d10	-1,858	d22	-2,162	d10	-1,786	d10	-1,864	d10	-1,687
d22	-2,090	d17	-2,336	d22	-2,579	d22	-2,109	d22	-2,235
d11	-2,606	d11	-2,472	d11	-2,788	d8	-2,258	d11	-2,595
d8	-2,676	d8	-2,624	d8	-2,949	d7	-2,296	d8	-2,627
d7	-2,715	d7	-3,399	d7	-3,062	d11	-2,513	d7	-2,868

Tab. 7: Popolazione degli studenti frequentanti Architettura, anni 2005, 2004 e 2003: test Chi-quadrato per il confronto della soddisfazione e per la significatività del DIF rispetto all’insegnamento.

Var.raggruppamento	Anno	Chi-quadro	GdL	Prob.
Insegnamento	2005	167,9	13	0,000
Item*Insegnamento	2005	1241,5	221	0,000
Insegnamento	2004	52,8	13	0,000
Item*Insegnamento	2004	890,6	221	0,000
Insegnamento	2003	162,1	13	0,000
Item*Insegnamento	2003	856,5	221	0,000

A questo punto è possibile stilare una graduatoria di soddisfazione per ciascun insegnamento di Architettura e per ognuno dei tre anni (cfr. Tab.

8) ed effettuare un confronto con le graduatorie che si ottengono da una semplice aggregazione delle risposte fornite al questionario. In particolare, in Tab. 9 vengono presentate le graduatorie ottenute dalla percentuale media di giudizi positivi alle domande del questionario prese in considerazione, mentre le graduatorie di Tab. 10 sono ricavate dalle risposte positive alla domanda sulla soddisfazione globale (d23). Dal confronto dei tre tipi di graduatorie ottenute emergono differenze piuttosto consistenti che inducono a riflettere sull'utilizzo diffuso delle graduatorie basate sulla semplice aggregazione delle risposte "grezze" ad un questionario. Inoltre, il ricorso alla domanda globale presenta l'ulteriore svantaggio di non consentire una chiara discriminazione tra insegnamenti quando, come nel caso considerato, la distribuzione di soddisfazione presenta una asimmetria accentuata (in tal caso verso l'alto): si veda, in particolare, la graduatoria del 2004, in cui ben 6 insegnamenti su 14 hanno ottenuto il 100% di risposte positive.

La presenza di DIF a livello di insegnamento e la variabilità osservata tra le graduatorie di soddisfazione nel corso negli anni, induce ad approfondire l'analisi in tale direzione. A scopo esemplificativo, è stato preso in esame l'insegnamento 20593 (Laboratorio di restauro), il quale mostra un livello di soddisfazione media decrescente (cfr. Tab. 8) - +0,138 logit nel 2003, +0,099 logit nel 2004 e -0,121 logit nel 2005 - che si traduce in una perdita di 5 posizioni in graduatoria dal 2004 al 2005, mentre dal 2003 al 2004 non si notano cambiamenti. Ad un esame più attento (cfr. Tab. 11) si osserva che le differenze nel livello di soddisfazione non sono significative (probabilità del Chi-quadrato pari a 0,1585), mentre risulta una presenza significativa di DIF rispetto all'anno di rilevazione. I due risultati sono solo apparentemente contrastanti: infatti il DIF è inerente le stime di difficoltà degli item e la sua presenza indica che nel corso dei tre anni esaminati ci sono state modifiche sostanziali nella percezione della difficoltà di uno o più item che hanno avuto effetto negativo ma non significativo sul livello di soddisfazione. In effetti, esaminando le graduatorie di difficoltà dei tre

Tab. 8: Popolazione degli studenti frequentanti Architettura, anni 2005, 2004 e 2003: graduatoria di soddisfazione degli insegnamenti; *Rasch model.*

	2005		2004		2003	
Insegnamento	Soddisf.	Posizione	Soddisf.	Posizione	Soddisf.	Posizione
29101	0,507	1	0,016	6	0,435	1
20509	0,305	2	-0,175	13	0,067	8
29341	0,214	3	-0,168	12	0,104	6
29173	0,196	4	-0,014	8	0,255	3
29111	0,182	5	0,147	3	-0,166	11
25005	0,117	6	0,009	7	0,225	4
29083	0,054	7	0,420	1	0,021	10
25001	-0,053	8	-0,302	14	-0,533	12
20593	-0,121	9	0,099	4	0,138	5
29117	-0,140	10	-0,163	11	-0,539	13
29236	-0,148	11	0,169	2	0,391	2
29250	-0,260	12	-0,033	9	0,049	9
29181	-0,279	13	-0,076	10	-0,545	14
29009	-0,576	14	0,071	5	0,098	7

Tab. 9: Popolazione degli studenti frequentanti Architettura, anni 2005, 2004 e 2003: graduatoria di soddisfazione degli insegnamenti- *Percentuale media di giudizi positivi (decisamente si o più si che no) alle 17 domande del questionario prese in considerazione.*

	2005		2004		2003	
Insegn.	Soddisf.	Posizione	Soddisf.	Posizione	Soddisf.	Posizione
29236	94,1	1	97,8	1	89,9	1
29101	86,4	2	86,9	3	84,5	3
29111	85,6	3	92,3	2	87,1	2
25001	84,9	4	73,5	14	69,1	13
20509	83,2	5	75,2	13	79,0	9
29173	83,1	6	79,6	7	81,3	6
25005	81,3	7	79,1	8	80,2	7
29341	79,2	8	76,5	11	81,9	5
20593	78,2	9	80,7	5	79,9	8
29083	76,6	10	85,5	4	77,5	11
29117	76,5	11	75,4	12	70,6	12
29181	70,8	12	76,9	10	65,5	14
29250	69,9	13	80,1	6	82,3	4
29009	64,3	14	78,1	9	78,7	10
Tot.	78,3		79,0		77,9	

Tab. 10: Popolazione degli studenti frequentanti Architettura, anni 2005, 2004 e 2003: graduatoria di soddisfazione degli insegnamenti - *Percentuale di giudizi positivi (decisamente si o più si che no) alla domanda sulla soddisfazione globale per l'insegnamento (d23).*

	2005			2004			2003		
Insegn.	Soddisf.	Posizione	N°	Soddisf.	Posizione	N°	Soddisf.	Posizione	N°
29236	100,0	1	5	100,0	1	8	100,0	1	7
29111	100,0	1	9	100,0	1	13	100,0	1	10
25001	100,0	1	14	85,6	12	90	74,3	13	74
29101	93,9	4	33	100,0	1	13	92,9	4	42
29341	92,1	5	38	88,9	10	18	84,6	10	13
20509	92,0	6	88	89,5	9	38	90,8	6	152
29173	87,5	7	24	100,0	1	15	91,4	5	35
20593	83,3	8	48	93,0	7	100	86,6	9	97
29083	82,5	9	40	100,0	1	39	82,0	11	61
25005	82,5	9	40	91,3	8	80	89,7	7	58
29181	80,0	11	25	79,3	13	29	57,1	14	14
29117	76,5	12	17	75,0	14	32	76,1	12	46
29009	70,9	13	55	87,2	11	39	88,9	8	18
29250	62,5	14	16	100,0	1	8	100,0	1	14
Tot.	85,0		452	90,0		522	86,0		641

Tab. 11: Popolazione degli studenti frequentanti l'insegnamento 20593 di Architettura, anni 2005, 2004 e 2003: test Chi-quadrato per il confronto della soddisfazione e per la significatività del DIF rispetto all'anno di frequenza.

Var.raggruppamento	Chi-quadro	GdL	Prob.
Anno	3,7	2	0,1585
Item*Anno	67,8	34	0,0005

anni (cfr. Tab. 12) si notano variazioni nelle posizioni occupate da parte di diversi item e nelle relative stime di difficoltà: a fronte di una costante riduzione di difficoltà da parte delle domande d1, d3 e d20, si ha un aumento di difficoltà per le questioni d2, d6, d8, mentre la d7 e la d12 presentano un andamento variabile ma meno chiaro.

Naturalmente, l'interpretazione di questi risultati deve essere demandata a chi ha potere decisionale e può essere ritenuto capace di influenzare almeno in parte le percezioni e le opinioni degli studenti: per quanto riguarda le domande inerenti il singolo insegnamento si tratterà del docente

Tab. 12: Popolazione degli studenti frequentanti l'insegnamento 20593, anni 2005, 2004 e 2003: graduatoria di difficoltà degli item, per anno di frequenza. (*= statistiche Outfit e/o Infit non significative).

	2005			2004			2003	
	Item	Difficoltà		Item	Difficoltà		Item	Difficoltà
1	d2	0,232	1	d17*	-0,009	1	d17	-0,140
2	d6	-0,250	2	d20	-0,291	2	d20	-0,191
3	d17*	-0,443	2	d21*	-0,291	3	d21*	-0,662
3	d21	-0,443	4	d2	-0,843	4	d2	-0,886
5	d4	-0,739	5	d6	-0,897	4	d1	-0,887
6	d20	-1,168	6	d1	-1,066	6	d6	-1,066
7	d1	-1,815	7	d19	-1,124	7	d4	-1,322
7	d19	-1,815	8	d4	-1,373	7	d19	-1,609
9	d9	-1,973	9	d9	-1,733	9	d3	-1,853
10	d7	-2,146	10	d14	-2,290	10	d9	-2,348
10	d14	-2,146	11	d22	-2,671	10	d22	-2,348
12	d22	-2,712	12	d10	-2,826	12	d14	-2,468
13	d8	-2,824	13	d3	-3,206	13	d7	-2,598
14	d10	-3,151	14	d12	-3,452	13	d11	-2,598
15	d3	-3,596	15	d7	-3,763	15	d10	-2,900
15	d11	-3,596	15	d8	-3,763	16	d8	-3,287
17	d12	-3,597	17	d11	-4,192	17	d12	-3,850

titolare della cattedra, mentre relativamente alle domande sul corso di studi la possibilità di individuare opportuni interventi correttivi è di pertinenza del consiglio di corso di laurea.

Diversamente da tutti i modelli di Rasch stimati precedentemente, ciascuno dei modelli relativi ai tre anni presenta un buon adattamento complessivo ai dati, eccezion fatta per le domande d17 per il 2004 e 2005 e d21 per il 2003 e 2004, che evidentemente contribuiscono a misurare un costrutto latente diverso dagli altri item. Questo risultato induce a pensare che il singolo insegnamento costituisca un livello di aggregazione degli studenti omogeneo tale da consentire un buon funzionamento del questionario della valutazione della didattica.

Infine, un'ulteriore fonte di riflessione è rappresentata dall'analisi con-

giunta della distribuzione della soddisfazione per gli studenti del medesimo insegnamento e della distribuzione della difficoltà degli item. Con riferimento alla Fig. 1, si osserva che la dicotomizzazione delle categorie di risposta (positive verso negative) non consente al questionario di discriminare in modo soddisfacente tra i vari livelli di soddisfazione di individui diversi, in quanto la distribuzione della soddisfazione (avente media pari a 0) è notevolmente disallineata verso l'alto rispetto alla distribuzione di difficoltà. A fini pratici, ciò significa che, al di là delle differenze che emergono tra i vari item, in generale nella popolazione analizzata è molto probabile individuare studenti soddisfatti in qualche misura degli aspetti inerenti la didattica indagati dal questionario (quindi studenti che nelle varie domande scelgono le modalità "più sì che no" o "decisamente sì"). Se si ritiene che gli item inseriti nel questionario siano esaustivi degli aspetti rispetto a cui interessa indagare sull'opinione degli studenti, il risultato è positivo, perché indice di una generale soddisfazione. D'altra parte, i risultati derivanti dalla stima del Partial Credit Model (cfr. Fig. 2), mantenendo distinte le quattro categorie di risposta, mostrano un quadro leggermente diverso: le due distribuzioni sono molto meno disallineate rispetto al caso precedente, tanto che adesso è possibile individuare soglie di alcuni item (la terza soglia degli item d17, d2, d4, d20, d21) il cui "superamento" è molto improbabile per qualsiasi individuo. Si può, quindi, concludere[10] che se da una parte la popolazione analizzata è generalmente soddisfatta, dall'altra la soddisfazione raggiunge comunque livelli elevati con minore probabilità: in altri termini, è mediamente probabile per un soggetto scegliere la modalità di risposta "più sì che no", ma è molto meno probabile la scelta della modalità "decisamente sì". A conferma di ciò, la mappa delle distribuzioni mostra che la seconda soglia di quasi tutti gli item si posiziona al di sotto del livello di abilità media (cioè 0). Inoltre, il fatto che in corrispondenza della prima

[10] *Anche le distribuzioni relative agli anni 2003 e 2004, non riportate nel testo, presentano un andamento simile a quelle relative al 2005.*

soglia (relativa alla scelta della modalità "più no che sì" rispetto a "decisamente no") di molti item non sia posizionato nessun individuo avvalora la tesi di una popolazione di studenti sostanzialmente soddisfatta dell'insegnamento analizzato, anche se non si raggiungono complessivamente livelli troppo elevati.

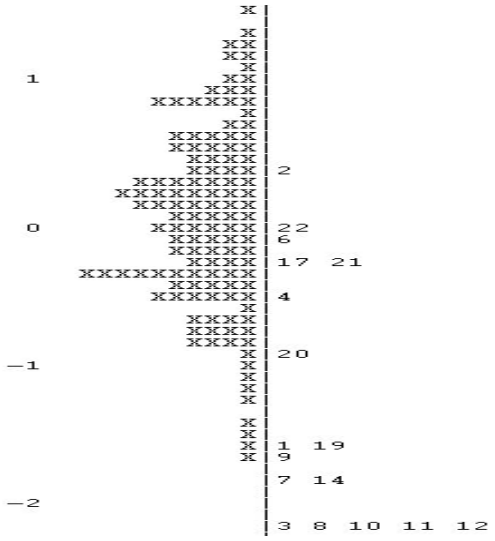

Fig. 1: Popolazione degli studenti frequentanti l'insegnamento 20593, anno 2005: mappa della distribuzione di soddisfazione e della distribuzione di difficoltà (X = 0,4 individui) - *Modello di Rasch Dicotomico.*

4. CONCLUSIONI E SVILUPPI FUTURI

L'analisi sviluppata nel paragrafo precedente ha posto in evidenza le potenzialità del modello di Rasch quale strumento per la valutazione delle performance e per la programmazione di interventi migliorativi del sistema universitario. Contemporaneamente, è stato possibile sottolineare come l'applicazione di tale modello a sistemi complessi, quale appunto quello universitario, richieda un'attenzione particolare a causa della natura fortemente disomogenea della popolazione presa in considerazione. Ignorare

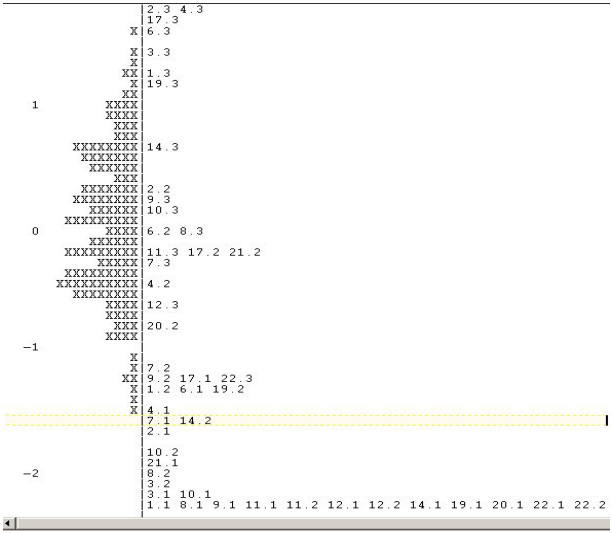

Fig. 2: Popolazione degli studenti frequentanti l'insegnamento 20593, anno 2005: mappa della distribuzione di soddisfazione e della distribuzione di difficoltà (X = 0,3 individui) - *Partial Credit Model.*

questa disomogeneità porta a stimare un modello che mal si adatta ai dati osservati e che, quindi, fornisce informazioni (sugli elementi critici su cui agire, ad es.) fuorvianti o, comunque, incomplete. La soluzione adottata in questa fase iniziale dell'analisi dei dati dell'Ateneo fiorentino è molto semplice e si basa sulla stima di modelli di Rasch separati per ciascun gruppo omogeneo di studenti, laddove i gruppi omogenei sono stati individuati nei singoli insegnamenti. L'approccio adottato è piuttosto empirico e possibili approfondimenti e sviluppi sono senz'altro possibili.

In primo luogo, sarebbe utile una più approfondita analisi del concetto di gruppo omogeneo, poiché non necessariamente il singolo insegnamento è il livello minimo di aggregazione, ma in certi casi due o più insegnamenti potrebbero essere raggruppati insieme (potrebbe essere il caso di corsi tenuti dal medesimo docente oppure relativi allo stesso corso di studi e aventi caratteristiche simili). A questo proposito potrebbero essere utilmente impiegate tecniche di statistica multivariata, tra cui la *cluster analysis*.

L'altro aspetto di una certa rilevanza concerne la possibilità di considerare simultaneamente le diverse popolazioni di studenti, ricorrendo ad un modello di Rasch opportunamente esteso. Nel caso specifico, la natura complessa di un sistema universitario si sostanzia in una struttura di tipo multilivello, dove le unità di primo livello sono rappresentate dagli studenti, quelle di secondo livello dagli insegnamenti, quelle di terzo livello dai corsi di laurea, che a loro volta sono aggregati in facoltà e in tipologie diverse (lauree di primo livello, di secondo livello, specialistiche a ciclo unico e lauree pre-riforma)[11]. Una soluzione metodologica per evitare l'oneroso lavoro di stimare modelli di Rasch separati per ciascun insegnamento e per tenere in debita considerazione la struttura gerarchica dei dati, è dunque quella di adattare **modelli di Rasch multilivello**. In appendice B viene presentata una breve descrizione di questa tipologia di modelli, con l'intento di far poi seguire un'analisi empirica dei dati sulla valutazione della didattica.

Infine, un ulteriore aspetto da prendere in considerazione per futuri sviluppi del lavoro riguarda il trattamento dei dati mancanti. Nell'analisi svolta fino a questo momento le risposte mancanti sono state ignorate in fase di stima dei modelli, mentre sarebbe opportuno valutare la possibilità di adottare idonei metodi di imputazione o, comunque, di metodi che consentano di tenere esplicitamente in considerazione la presenza di risposte mancanti non casuali; si veda a questo proposito il lavoro di (Wang, Chen & Sheu 2006) per un'estensione del modello di Rasch al caso di *missing* informativo con implementazione in SAS tramite la procedura Proc Nlmixed.

---

[11] *Un'ulteriore fonte di complessità è data dal fatto che gli stessi studenti frequentano più insegnamenti e, di conseguenza, le risposte ai relativi questionari non possono essere considerate indipendenti. Purtroppo, essendo i questionari anonimi, non è possibile tenere conto di questo elemento.*

A. IL QUESTIONARIO PER LA VALUTSZIONE DELLA DIDATTICA

Di seguito sono riportate le domande del questionario sulla valutazione della didattica utilizzato dall'Università di Firenze ed analogo (salvo poche modifiche) a quello proposto in sede di Comitato Nazionale di Valutazione del Sistema Universitario (R. di R. 01/00 e Doc. 09/02); si tratta di 23 domande con 4 modalità di risposta ordinali: decisamente no, più no che sì, più sì che no, decisamente sì. Oltre a queste domande, vengono acquisite informazioni su alcune caratteristiche dei rispondenti: maturità conseguita, se si tratta di studente a tempo pieno o parziale, tipologia di iscrizione, anno di corso, frequenza del corso, numero di studenti che hanno frequentato il corso nel periodo di riferimento. Inoltre, vengono proposti una serie di suggerimenti in forma chiusa relativi all'insegnamento e viene lasciato uno spazio per eventuali osservazioni personali.

- Organizzazione del corso di studi
  - d1. Il carico di lavoro complessivo degli insegnamenti ufficialmente previsti nel periodo di riferimento (bimestre, trimestre, semestre, ecc.) è accettabile?
  - d2. L'organizzazione complessiva (orario, esami intermedi e finali) degli insegnamenti ufficialmente previsti nel periodo di riferimento (bimestre, trimestre, semestre, ecc.) è accettabile?
- Organizzazione dell'insegnamento:
  - d3. Il carico di studio di questo insegnamento è proporzionato ai crediti assegnati?
  - d4. Il materiale didattico (assegnato o fornito) è adeguato per lo studio della materia?
  - d5. Le attività didattiche integrative (esercitazioni, laboratori, seminari, ecc.) risultano utili ai fini del'apprendimento?

  - d6. Le modalità di esame sono state definite in modo chiaro?

- Aspetti relativi alla docenza:
  - d7. Gli orari di svolgimento dell'attività didattica sono rispettati?
  - d8. Il personale docente è effettivamente reperibile per chiarimenti e spiegazioni?
  - d9. Il docente stimola/motiva l'interesse verso la disciplina?
  - d10. Il docente espone gli argomenti in modo chiaro?
  - d11. Il docente è disponibile ed esauriente in occasione di richieste di chiarimento?

- Aspetti specifici del corso di studi (*domande definite dalle singole facoltà*)
  - d12.
  - d13.
  - d14.
  - d15.
  - d16.

- Aule ed attrezzature
  - d17. Le aule in cui si svolgono le lezioni sono adeguate (si vede, si sente, si trova posto)?
  - d18. I locali e le attrezzature per le attività didattiche integrative (esercitazioni, laboratori, seminari, ecc.) sono adeguati?

- Informazioni aggiuntive e soddisfazione
  - d19. Le conoscenze preliminari possedute sono risultate sufficienti per la comprensione degli argomenti trattati?

- d20. La frequenza alle lezioni e/o esercitazioni è accompagnata da una regolare attività di studio?
- d21. Gli argomenti trattati sono risultati nuovi rispetto a quelli affrontati in insegnamenti precedenti?
- d22. Sei interessato agli argomenti dell'insegnamento?
- d23. Sei complessivamente soddisfatto dell'insegnamento?

B. IL MODELLO DI RASCH MUTILIVELLO

Dal punto di vista teorico, il modello di Rasch multilivello non presenta complicazioni particolari rispetto ad un qualsiasi modello multilivello. L'approccio più semplice è quello proposto da Kamata (2006), il quale unisce la struttura multilivello relativa alla natura gerarchica dei dati con la struttura multilivello propria dei modelli di Rasch.

Si prenda in considerazione il modello dicotomico dell'equazione 1[12]. Tale modello può essere interpretato in termini di modello a due livelli con intercetta casuale: le risposte agli item costituiscono le unità di primo livello, mentre gli studenti sono le unità di secondo livello. In particolare, il modello di I livello può essere scritto come:

$$logit(P_{ij}) = \log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{i0} + \beta_{i1}X_{i1} + \ldots + \beta_{i(J-1)}X_{i(J-1)} \qquad (6)$$

dove: i indica la generica unità di II livello (lo studente), j indica l'item ($j = 1, 2, \ldots, J-1$), P_{ij} è la probabilità che il soggetto i-esimo scelga la modalità di risposta 1 all'item j e X_{ij} è una variabile *dummy* che assume valore 1 quando l'osservazione è sul j-esimo item e valore 0 altrimenti.

Il modello di II livello è invece dato da:

$$\begin{cases} \beta_{i0} & = & \gamma_{00} + u_{i0} \\ \beta_{i1} & = & \gamma_{01} \\ \vdots & & \\ \beta_{i(J-1)} & = & \gamma_{0(J-1)} \end{cases} \qquad (7)$$

dove $u_{i0} \sim N(0, \tau)$. Quindi, β_{i0} è il termine di intercetta costituito da una componente fissa (γ_{00}) e da una componente casuale (u_{i0}) di II livello, mentre $\beta_{ij} = \gamma_{0j}$ è il coefficiente fisso associato con la *dummy* X_{ij}.

Mettendo insieme le due equazioni, si ottiene un modello a due livelli con intercetta casuale che è identico al modello di Rasch dicotomico del-

[12] *Nel caso di un Partial Credit Model sarà necessario estendere opportunamente il modello.*

l'equazione 1 con $x_{ij} = 1$ (basta semplificare l'equazione 1 portando il numeratore al denominatore):

$$P_{ij} = P(X_{ij} = 1) = \frac{1}{1 + \exp\{-[u_{i0} - (-\gamma_{0j} - \gamma_{00})]\}} \tag{8}$$

dove u_{i0} è l'abilità dell'individuo i-esimo (indicata con θ_i nell'equazione 1) e $(-\gamma_{0j} - \gamma_{00})$ è la difficoltà dell'item j-esimo (indicata con β_j nell'equazione 1), mentre $-\gamma_{00}$ è la difficoltà dell'item di riferimento (il J-esimo).

L'estensione al caso multilivello consiste nell'aggiungere un livello per ogni grado di aggregazione. Ipotizzando di essere interessati soltanto a sviluppare una struttura a due livelli, in cui le unità di I livello sono gli studenti e quelle di II livello sono gli insegnamenti ($m = 1, 2, \ldots, M$), il modello di Rasch multilivello che ne risulta è un modello a tre livelli di aggregazione:

- Modello di I livello (*modello a livello di item*): è uguale al modello dell'equazione 6 con l'aggiunta del pedice m ad indicare il terzo livello di aggregazione.

  $$logit(P_{ijm}) = \log\left(\frac{P_{ijm}}{1 - P_{ijm}}\right) = \beta_{i0m} + \beta_{i1m}X_{i1m} + \ldots + \beta_{i(J-1)m}X_{i(J-1} \tag{9}$$

  dove X_{ijm} è la j-esima variabile dummy per lo studente i-esimo che frequenta l'insegnamento m.

- Modello di II livello (*modello a livello di studente*): anche questo è identico al modello dell'equazione 7 con l'eccezione dell'aggiunta del pedice m:

  $$\begin{cases} \beta_{i0m} & = & \gamma_{00m} + u_{i0m} \\ \beta_{i1m} & = & \gamma_{01m} \\ \vdots & & \\ \beta_{i(J-1)m} & = & \gamma_{0(J-1)m} \end{cases} \tag{10}$$

con $u_{i0m} \sim N(r_{00m}, \tau_\gamma)$ che indica quanto la soddisfazione dello studente i-esimo per l'insegnamento m-esimo si discosta dalla soddisfazione media r_{00m} per l'insegnamento m-esimo. Per ipotesi, la varianza τ_γ è assunta identica per tutti gli insegnamenti.

- Modello di III livello (*modello a livello di insegnamenti*):

$$\begin{cases} \gamma_{00m} & = & \pi_{000} + r_{00m} \\ \gamma_{01m} & = & \pi_{010} \\ \vdots & & \\ \gamma_{0(J-1)m} & = & \pi_{0(J-1)0} \end{cases} \tag{11}$$

con $r_{00m} \sim N(0, \tau_\pi)$.

Si osservi che nella versione di Kamata i coefficienti da γ_{01m} a $\gamma_{0(J-1)m}$, che indicano la difficoltà degli item, hanno soltanto una componente fissa, cioè sono costanti tra le unità di III livello: dai risultati ottenuti nell'analisi svolta al precedente paragrafo emerge, invece, che, a causa della presenza di DIF in relazione alla variabile "insegnamento", è necessario prevedere livelli di difficoltà diversi per ogni insegnamento e, quindi, coefficienti $\gamma_{01m}, \ldots, \gamma_{0(J-1)m}$ casuali. Se a livello teorico questo non crea problemi di alcun tipo (basta aggiungere una componente casuale ad ognuna delle equazioni in 11), a livello di stima il modello, già di per sé complesso, si complica notevolmente, a causa dell'incremento nel numero di componenti di varianza e covarianza da stimare. Una soluzione può essere quella di condurre analisi esplorative in modo da capire quali item sono significativamente distorti in relazione all'insegnamento e quali invece mostrano un funzionamento costante: soltanto per i primi sarà necessario prevedere una componente casuale. Con riferimento all'esempio sviluppato nel precedente paragrafo (cfr. Tab. 6), ad esempio, le domande d7, d8, d11, d1 e d20 non presentano problemi di DIF rispetto alla variabile "corso di laurea", al contrario delle d17, d19 e d6.

L'unione dei tre modelli dà origine alla seguente equazione, analoga all'equazione 8:

$$P_{ijm} = P(X_{ijm} = 1) = \frac{1}{1 + \exp\{-[(r_{00m} + u_{i0m}) - (-\pi_{0j0} - \pi_{000})]\}} \quad (12)$$

dove:

$r_{00m} + u_{i0m}$ = soddisfazione dello studente i-esimo per l'insegnamento m-esimo;

r_{00m} = soddisfazione media degli studenti per l'insegnamento m-esimo;

u_{i0m} = componente specifica dello studente nell'insegnamento m; indica quanto la soddisfazione specifica dello studente i-esimo devia rispetto al valore medio dell'insegnamento.

$-\pi_{0j0} - \pi_{000}$ = difficoltà dell'item j-esimo, definita come scostamento rispetto alla difficoltà π_{000} dell'item di riferimento J.

La logica seguita per presentare il modello di Rasch a due livelli (equivalente ad un modello a tre livelli) può essere facilmente estesa al caso in cui si vogliano considerare ulteriori livelli di aggregazione (corsi di laurea, facoltà, ecc.) e dati longitudinali[13]. I maggiori problemi sorgono in fase di implementazione: i software specifici per i modelli di Rasch non prevedono la possibilità di estensioni al caso multilivello, mentre software statistici più generici (ad es. la routine Gllamm di Stata), che godono di una maggiore flessibilità, incontrano gli ovvi problemi di stima derivanti dalla complessità di questa tipologia di modelli, dovuta essenzialmente all'elevato numero di effetti casuali. Quindi, ulteriori approfondimenti sono ancora necessari, come è testimoniato dalla scarsità di applicazioni presenti in letteratura (si vedano per alcuni esempi Skrondal & Rabe-Hesketh (2002) e Pastor & Beretvas (2006)).

---

[13] *L'analisi dei dati longitudinali può essere affrontata nella logica multilivello, essendo le misure ripetute le unità di primo livello e i soggetti misurati le unità di secondo livello, di conseguenza l'estensione al modello di Rasch è ottenibile inserendo un livello ulteriore nel modello multilivello.*

RIFERIMENTI BIBLIOGRAFICI

BAKER, F. & KIM, S. (2004), *Item response theory. Parameter estimation techniques*, Dekker.

BINI, M. & CHIANDOTTO, B. (2003), 'La valutazione del sistema universitario italiano alla luce della riforma dei cicli e degli ordinamenti didattici', *Studi e Note di Economia* **2**, 29–61.

BOND, T. & FOX, C. (2001), *Applying the Rasch model: fundamental measurement in the human sciences*, Lawrence Erlbaum Associates.

CHIANDOTTO, B. (2002), Valutazione dei processi formativi: cosa, come e perchè, *in* M. D'Esposito, ed., 'Valutazione della Didattica e dei Servizi nel Sistema Università', Salerno: CUSL.

CHIANDOTTO, B. (2004), 'Sulla misura della qualità della formazione universitaria', *Studi e note di economia* **3**, 27–61.

FISCHER, G. (1995), Derivations of the rasch model, *in* G. H. Fischer & I. W. Molenaar, eds, 'Rasch models. Foundations, recent developments, and applications.', Springer-Verlag, pp. 15–38.

GLAS, A. & VERHELST, N. (1995), Tests of fit for polytomous rasch models, *in* G. H. Fischer & I. W. Molenaar, eds, 'Rasch models. Foundations, recent developments, and applications.', Springer-Verlag, pp. 325–352.

GORI, E., SANARICO, M. & PLAZZI, G. (2005), 'La valutazione e la misuraizone nelle scienze sociali: oggettività specifica, statistiche sucienti e modello di rasch', *Non Profit* **3**, 605–644.

GORI, E. & VITTADINI, G. (1999), La valutazione dell'ecienza ed ecacia dei servizi alla persona. impostazione e metodi., *in* E. Gori & G. Vittadini, eds, 'Qualità e valutazione nei servizi di pubblica utilità', ETAS, pp. 121–241.

KAMATA, A. (2006), 'Procedure to perform item response analysis by hierarchical generalized linear model', In press on Florida *Journal of Educational Research.*

MOLENAAR, I. (1995), Estimation of item parameters, *in* G. H. Fischer & I. W. Molenaar, eds, 'Rasch models. Foundations, recent developments, and applications.', Springer-Verlag, pp. 39–51.

PASTOR, D. & BERETVAS, S. (2006), 'An illustration of longitudinal rasch modeling in the context of psychotherapy outcomes assessment', In press on Applied Psychological Measurement.

SKRONDAL, A. & RABE-HESKETH, S. (2002), Generalized Latent Variable Modeling. Multilevel, Longitudinal, and Structural Equation Models, Chapman and Hall.

TESIO, L., VALSECCHI, M., SALA, M., GUZZON, P. & BATTAGLIA, M. (2002), 'Level of activity in profound/severe mental retardation (lapmer): a raschderived scale of disability', *Journal of Applied Measurement* **3(1)**, 50–84.

WANG, W., CHEN, C. & SHEU, C. (2006), 'Formulating multidimensional item response models using the sas nlmixed procedure', In http://inoce.adm.ccu.edu.tw/edu/93paperCCT.doc.

WRIGHT, B. & MASTERS, G. (1982), *Rating scale analysis*, Mesa Press.

WU, M., ADAMS, R. & WILSON, M. (1998), Acer Conquest. Generalised item response modelling software, Acer Press.

RASCH MODELS AND EVALUATION OF THE UNIVERSITY DIDACTICS

Summary

This paper concerns with evaluation of quality of services from complex systems, such as the university one. Particularly, the interest is facing to measure the satisfaction of the attending students for the university didactics. Because of the latent nature of the studied variable, it is necessary to define statistical instruments to measure the satisfaction objectively, through a synthesis of responses to the items of an ad hoc questionnaire by attending students. To such purpose the potentiality of Rasch models are analyzed, as reference method to the evaluation of complex systems. The empirical analysis has been conducted on data collected at the University of Florence in the years 2003, 2004 and 2005.