***S**tatistica & **A**pplicazioni*
Vol. III, n. speciale 1, 2005
pp. 69-76

Dalle valutazioni soggettive alle misure oggettive: applicazioni del modello Multifacet

Enrico Gori[§]

Guido Gay[‡]

Summary: *This paper deals with the problem of the treatment of ordinal data arising from the evaluation of statistical units (projects) expressed by judges with respect a set of predefined common criteria. The methodology proposed to create objective interval scaled measures is the Rasch Multifacet modes. The application regards a set of project, that was selected for financing by the Lombardia regional Government. We found that the scale used to evaluate the projects contains to many modalities: as we aggregate this modalities to get a simpler scale most of misfitting observations disappeared and the criteria used to evaluate the projects look pretty good in term of fit and reliability. Some judge show very lenient and other were, on the contrary, very restrictive in their judgement: the model adjusted adequately for these differences, producing ranking based on the estimated measures that were quite different from that obtained from the observed scores.*

Keywords: *ordinal measures, interval scale objective measures, judges, Rasch Multifacet model*

1. Introduzione

In molti campi – istruzione, ricerca, sport, finanza, medicina - le valutazioni vengono spesso effettuate da esperti che devono esprimere giudizi relativamente ad aspetti difficilmente misurabili, detti "tratti latenti". Poiché tali valutazioni possono influenzare significativamente le prospettive future degli individui o degli enti sottoposti ad esame, è necessario ricercare la massima oggettività possibile nel processo di valutazione. Il grado di

[§] Dipartimento di Scienze Statistiche – Università degli Studi di Udine – via Treppo, , 33100 UDINE (e-mail: e.gori@dss.uniud.it).

[‡] Istituto Regionale di Ricerca della Lombardia – via G. Copernico , 38, 20100 MILANO (e-mail: gay@irer.it).

oggettività di tale processo dipende da 2 fattori fondamentali: a) la possibilità di ridurre al minimo la *soggettività* nella misura dei tratti latenti b) la *validità di contenuto*, ossia l'adeguatezza delle prove ai fini della valutazione del "tratto latente" che si desidera misurare.

La validità di contenuto dipende in modo essenziale dalle teorie sostanziali a monte del processo di misurazione, come argomentato con forza in (Bond, 2003): "I am now much more convinced that measurement in the human sciences must be theoretically driven. In common with other quantitative rational sciences, we neeed theories of measurement of human variables which satisfy the requirement for scientific measurement. On the other hand, we need substantive theories about the human condition that allow us to examine how the responses that candidates make to our data collection devices are connected with the human attribute under investigation. While I might have been fortunate to have Piaget's 60 books and 600 articles on one side, and advice to use the Rasch model to solve developmental measurement problems on the other, many attemps at scale-building across the human sciences have a distictly bottom-up approach".

Il problema della soggettività invece può essere affrontato utilizzando un modello matematico che realisticamente parte dal riconoscimento del fatto che la valutazione delle prove o delle attività, di un individuo o di un ente, da parte di giudici, è influenzata da tre fattori fondamentali: l'*abilità* del *soggetto* - ossia la bravura di un atleta, la bontà di un progetto, la rischiosità di un'azienda; la *difficoltà* delle diverse *prove*; la *severità* del *giudice*. La base di partenza per la costruzione di un tale modello è costituita dalla ricerca del matematico danese Georg Rasch (Rasch, 1977) il quale, nel porsi il problema di individuare ciò che caratterizza la superiorità delle scienze naturali rispetto a quelle umane, giunse alla conclusione che il concetto di "scienza" è legato alla possibilità di sviluppare metodi per trasformare osservazioni in misure, secondo regole che soddisfino il principio della *oggettività specifica.* In termini intuitivi tale principio si riferisce al fatto che i metodi di misurazione delle scienze naturali consentono di misurare caratteristiche specifiche di un *soggetto* senza che il processo di misurazione risulti influenzato da caratteristiche del *soggetto* diverse da quella di interesse, da altri *soggetti*, e da particolarità dello strumento utilizzato a tale scopo. Affinché il procedimento di misura possieda tale proprietà è necessario che la "manipolazione" delle osservazioni nel sistema di riferimento avvenga attraverso una tipologia di modelli matematici ben precisa (Gori, et al., 2005) che sono appunto detti "modelli di Rasch".

2. Il modello di Rasch

Tutti i modelli di questo tipo si fondano su tre assunti (Hambleton e Swaminathan, 1985):

A1. Unidimensionalità. Esiste una entità unidimensionale θ_n, detta *abilità latente*, associata ad un generico soggetto *n*, che determina la sua capacità di superare la *prova* a cui è sottoposto; le *prove* sono relative a tale dimensione unica e sono caratterizzate da una *difficoltà* δ_i i = 1,2,…,*I*; i *giudici* sono caratterizzati da un parametro γ_j, j = 1,2,…,*J* detto *severità.*

A2. Monotonicità. rappresenta il risultato dell'incontro tra il soggetto *n,* la prova *i* ed il giudice *j* costituisce una variabile casuale che soddisfa la condizione che $P(X_{nij} > t|\theta_n, \delta_i, \gamma_j)$ sia una funzione monotòna della *abilità* θ_n, per ogni *i* e ogni *t* (l'argomento della funzione). Soggetti con *abilità* più elevate hanno una maggiore probabilità di rispondere correttamente, di superare le prove o ricevere una valutazione elevata. Questo assunto consente di utilizzare il vettore delle osservazioni $\mathbf{X}_n = \{X_{n1j}, X_{n2j}, \ldots, X_{nIj}\}$ relativo alle reazioni del soggetto *n* alle diverse prove, come una serie di misure ripetute sullo stesso soggetto.

A3. Indipendenza locale.

$$P(\mathbf{X}_n|\theta_n, \delta_1, \delta_2 \cdots \delta_I, \gamma_1, \gamma_2, \ldots \gamma_J) = \prod_{i=1}^{I}\prod_{j=1}^{J} P(X_{ni}|\theta_n, \delta_i, \gamma_j),$$

ossia, condizionatamente alla abilità del *soggetto*, le reazioni alle diverse *prove* e diversi giudici sono indipendenti tra loro.

Nel caso in cui siano presenti dei giudici, che esprimono una valutazione sulle *prove* prodotte dai *soggetti*, il modello che soddisfa il principio della *oggettività specifica* prende il nome di modello *multifacet* (Linacre e Wright, 1997). La versione del modello utilizzata nella analisi empirica effettuata postula la seguente condizione per la variabile casuale X_{nij} che rappresenta la "reazione" che consegue dall'incontro tra il soggetto *n*, la *prova i* e il *giudice j:*

$$P(X_{nij}) : \ln\frac{P(X_{nij} = k)}{P(X_{nij} = k-1)} = \theta_n - \delta_i - \gamma_j - \tau_k$$

dove δ_i rappresentano le difficoltà medie dei singoli criteri, i τ_k sono interpretabili come la *difficoltà* aggiuntiva per raggiungere il livello *k,* ipotizzate identiche per ogni criterio (come nella versione del modello *rating scale* senza la presenza di giudici), mentre γ_j sono le severità dei giudici. Si evidenzia che nell'applicazione che segue non tutti i criteri utilizzati presentano lo stesso numero di livelli, e l'indice *k* varia in maniera diversa per gruppi di criteri. Questo significa che nella verosimiglianza ogni criterio

entra con il proprio numero di livelli, ma si è imposto che le soglie siano identiche per ciascun gruppo con lo stesso numero di livelli, come nel modello *rating scale*. Si rileva che, nei modelli *multifacet*, secondo l'impostazione della misurazione oggettiva dato da Rasch, le interazioni tra giudici e soggetti, o giudici e prove, non sono ammissibili in quanto fanno venire meno la proprietà dell'*oggettività specifica*: le interazioni – come classificate da Lynch e McNamara (1998) - tra giudici e soggetti evidenziano fenomeni di "favoritismo" o "avversione" rispetto ai soggetti valutati; quelle tra giudici e prove evidenziano una mancata condivisione da parte dei giudici dell'ordine di importanza delle prove. Qualora queste interazioni sussistano, la via per la produzione di misure oggettive non è quella di accrescere il numero di parametri (con il rischio di ottenere modelli non identificabili) ma piuttosto quella di sostituire i giudici e/o i criteri utilizzati, al fine di ottenere osservazioni che soddisfano la condizione (7), unica garanzia di oggettività. Qui si evidenzia la netta differenza tra l'approccio statistico al problema (la ricerca di modelli che si adattano ai dati) e l'approccio della misurazione oggettiva (la ricerca di osservazioni che soddisfino il principio di oggettività specifica).

3. Descrizione del problema e dei dati disponibili

L'applicazione che segue è relativa a 89 progetti di ricerca in campo agricolo, presentati da università e centri di ricerca nell'ambito del "Programma di ricerca in campo agricolo 2004-2006" della Regione Lombardia, che verranno analizzati attraverso il modello *multifacet*.

Il processo di selezione dei progetti pervenuti, descritto compiutamente in una nota di ricerca dell'IReR (IReR, 2005), ha comportato una fase di valutazione finalizzata a verificare in particolare la qualità tecnico-scientifica del progetto, la competenza e la capacità operativa a gestionale dei soggetti attuatori, la qualità del piano di sfruttamento e di trasferimento dei risultati, la congruità del piano finanziario.

La valutazione è stata effettuata assegnando ad ogni progetto dei voti relativi a criteri quali la "Chiarezza e concretezza degli obiettivi del progetto", la "Qualità scientifica della ricerca e livello di innovazione", la "Presenza di indicatori di risultato e loro coerenza", la "Qualità del programma di iniziative di informazione e di trasferimento dei risultati" oppure la "Competenza degli attuatori valutata in base ai curricula presentati". Questi criteri non discendono esplicitamente da teorie sostanziali nel senso richiesto da Bond (2003) ma piuttosto da una un'analisi preliminare dei criteri utilizzati in processi di valutazione similari e da un processo di condivisione con i proponenti dei progetti. Nonostante questa procedura di selezione distintamente sostanzialmente ad-hoc, i criteri

utilizzati sono risultati ben strutturati ai fini della costruzione della scala di valutazione.

I voti sono stati attribuiti indipendentemente da almeno 2 valutatori per progetto, di cui uno assume il ruolo di coordinatore e l'altro (o gli altri) di supporto.

3.1 Analisi preliminare dei giudizi

Per ogni progetto si è proceduto a raccogliere, a partire dal materiale messo a disposizione dall'Amministrazione regionale i voti indipendentemente attribuiti dai valutatori ai progetti in base ai criteri di valutazione. I voti sono stati ricodificati in una scala numerica intera 0-10. Un'analisi preliminare dei dati a disposizione relativi ai 44 valutatori, 89 progetti e 14 criteri rileva quanto segue. La figura 1 riporta il punteggio medio - che sarà definito "grezzo" per distinguerlo da quello derivante dal modello - ottenuto da ogni progetto da parte del complesso dei valutatori in tutti i criteri, nonchè quello medio minimo e massimo rispetto all'insieme dei valutatori per ogni singolo progetto.

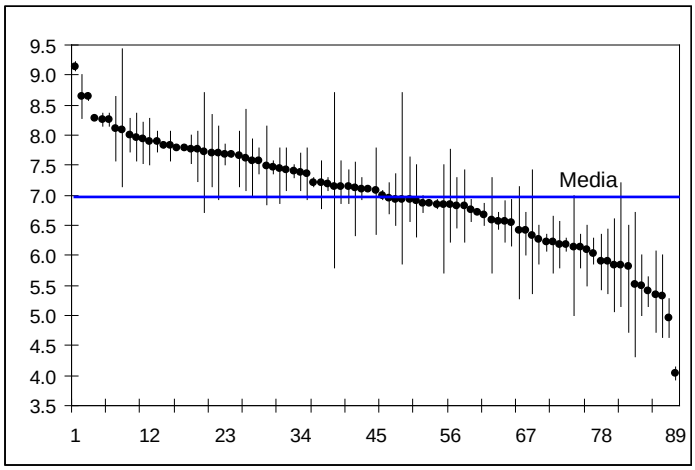

Figura 1. *Punteggio grezzo minimo, medio e massimo ottenuto dai progetti da parte dei valutatori*

L'analisi della distribuzione dei voti ha poi mostrato che non tutte le modalità sono state utilizzate e, in molti casi, sarebbe stato sufficiente adottare una scala con 2 o 3 modalità. Questo al fine di ridurre l'errore nell'attribuzione dei giudizi da parte dei valutatori.

Si è provveduto pertanto ad individuare, attraverso prove e tentativi successivi, la ricodifica delle scale di valutazione che consentisse una più netta separazione tra le curve di probabilità ed un migliore adattamento del

modello. Si è optato per una scala a tre modalità (del tipo insufficiente, sufficiente/buono, ottimo) tranne che in caso in cui è risultato più opportuno ridursi ad una valutazione dicotomica (inadeguato, adeguato).

Il modello finale stimato presenta, come documentato in (IReR, 2005), un ottimo adattamento al modello di Rasch e può essere dunque utilizzato per ottenere stime della bontà dei progetti, della severità dei giudici e della difficoltà dei criteri.

3.2 Analisi della distorsione indotta dall'uso dei punteggi grezzi

Sulla base di questi risultati è possibile avere diverse conferme dell'effetto distorsivo indotto dall'uso del punteggio grezzo al posto delle misure di bontà ottenute dal modello di Rasch. Dalla figura 2, si può osservare come il progetto 824, che presenta un rango relativo (ovvero il rango rapportato al valore massimo) pari a circa 0.30, con le misure di bontà, risulta invece assai meglio piazzato nella graduatoria sulla base del punteggio grezzo (oltre 0.70); viceversa il progetto 851, che presenta un rango relativo che supera 0.70 in base alle misure di bontà, è assai sfavorito sulla base del punteggio grezzo (appena 0.35). Come si può rilevare dalla figura, tali discrepanze sono dovute essenzialmente al fatto che il primo è valutato dai valutatori 3 e 26 che, nella scala di severità, si collocano in basso (sono quindi più generosi), mentre il secondo è valutato dai valutatori 14, 16 e 18 i quali si collocano invece nella parte alta di tale scala.

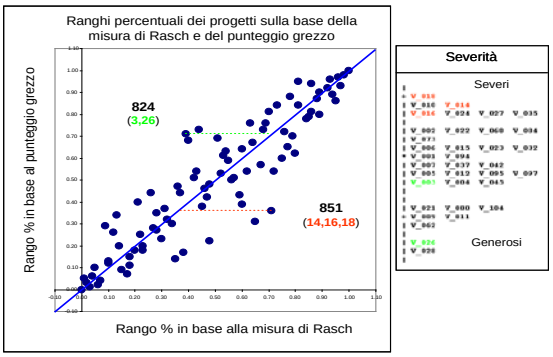

Figura 2 - *Effetto della diversa severità dei giudici sulla discrepanza nei ranghi percentuali*

4. Conclusioni

L'analisi qui condotta ha mostrato come i dati raccolti attraverso la valutazione dei progetti nell'ambito specificamente considerato, possano

essere adeguatamente utilizzati ai fini della costruzione di misure oggettive attraverso il modello *multifacet* con l'unica condizione di adottare scale di valutazione le quali, anziché spaziare da 0 a 10, siano ricondotte ad una semplice scala a tre modalità del tipo (insufficiente, sufficiente-buono, ottimo), che risulta in grado di ridurre l'eccessiva aleatorietà indotta dalla sovrabbondanza di modalità. Il modello così ottenuto presenta indici di infit e outfit del tutto accettabili, un buon adattamento globale e un'ottima affidabilità per la stima dei tre aspetti. Si segnala anche l'assenza di distorsione per le interazioni di cui al paragrafo 2, ed in particolare per quella tra valutatori e progetti che potrebbe nascondere fenomeni di favoritismo.

I criteri appaiono ben strutturati ai fini della costruzione della scala, e analisi successive con ulteriori dati potrebbero confermare la loro validità nella misura in cui le difficoltà stimate mostrassero una costanza nel tempo e nello spazio. Solo un paio di giudici presentano valori abnormi quanto a severità e, probabilmente, in un futuro potrebbero essere esclusi o adeguatamente formati, assieme a quelli che presentano indici di adattamento fuori della norma.

Complessivamente il modello *Multifacet* appare un utile strumento di analisi dei processi di valutazione basati su giudizi di esperti, una situazione molto diffusa nella pubblica amministrazione italiana.

Riferimenti Bibliografici

Andersen E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, **42**, 69-81.

Andrich D. (1978a). A rating scale formulation for ordered response categories. *Psychometrika*, **43**, 561-573.

Andrich D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, **38**, 665-680.

Andrich D. (1978c). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, **2**, 581-594.

Barndorff-Nielsen O. (1978). *Information and exponential families in statistical theory*. New York: J. Wiley.

Bond T. G., (2003). Validity and assessment: a Rasch measurement perspective. *Metodología de las Ciencias del Comportamiento* **5**, 179–194.

Cronbach L. J. (1949). #*Essentials of Psychological Testing*#. New York: Harper & Row. Quotations from the 1970 edition.

Embretson S. E. & Hershberger S. L. (1999). *The new rules of measurement: What every psychologist and educator should know. Mahwah,* NJ: Lawrence Erlbaum Associates.

Gori, E., Sanarico M., Plazzi G. (2005). La valutazione e la misurazione nelle scienze sociali: oggettività specifica, statistiche sufficienti e modello di Rasch. *Non Profit*, **3**.

Hambleton, R.K., Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer-Nijhoff.

IReR (2005). Analisi delle attività di valutazione tecnico-scientifica dei progetti di ricerca in campo agricolo attraverso il modello di Rasch

Karabatsos, G. (2001). The Rasch Model, Additive Conjoint Measurement, and New Models of Probabilistic Measurement Theory, *Journal of Applied Measurement*, **2**, 389–423.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1998). *A user's guide to FACETS: A Rasch measurement computer program*. Chicago: MESA Press.

Linacre, J.M., Wright, B.D. (1997). *FACETS: Many-Faceted Rasch Analysis*. Chicago: MESA Press

Lynch, B. K., McNamara, T. F. (1998). Using g-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, **15**, 158-80.

Masters, G. N., Keeves, J. P. (Eds.) (1999). *Advances in Measurement in Educational Research and Assessment*. New York: Pergamon (Elsevier Science).

Rasch, G. (1977). On Specific Objectivity: An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements. *Danish Yearbook of Philosophy*, **14**, 58-93. (http://www.rasch.org/memo18.htm)

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models. *Psychometrika*, **60**, 281-304.

Swaminathan, H. (1983). Parameter estimation in item response models. In R.Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia, 24-44.

Wilson, M., Engelhard, G.jr (Eds.) (2000). *Objective Measurement: Theory into Practice*-V. Stamford, CO: Ablex Publishing Corporation.

Wright, B.D., Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA Press.

Wright, B.D., Stone, M.H. (1979). *Best Test Design: Rasch Measurement*. Chicago: Mesa Press .

Wright, B., & Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement*, **1**, 83-106.