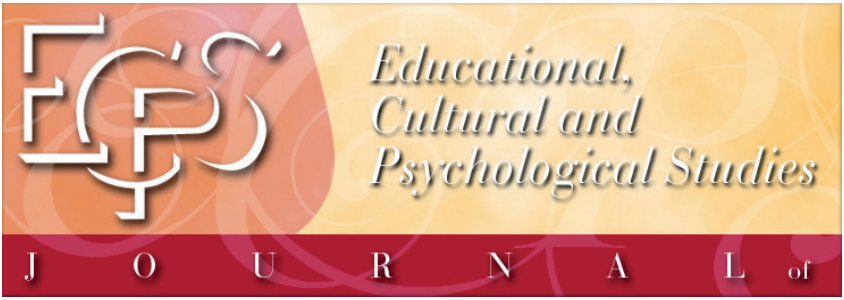

23
June 2021

ECPS Journal – 23/2021 - https://www.ledonline.it/ECPS-Journal/
Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-7916-976-9

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-7916-976-9

Recensioni
Reviews

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-7916-976-9

Accordo assoluto tra valutazioni espresse su scala ordinale

Giuseppe Bove - Daniela Marella

Università degli Studi Roma Tre - Department of Education (Italy)

DOI: https://dx.doi.org/10.7358/ecps-2021-023-boma

giuseppe.bove@uniroma3.it
daniela.marella@uniroma3.it

INTERRATER ABSOLUTE AGREEMENT FOR ORDINAL RATING SCALES

Abstract

Many methods for measuring agreement among raters have been proposed and applied in many domains in the areas of education, psychology, sociology, and medical research. A brief overview of the most used measures of interrater absolute agreements for ordinal rating scales is provided, and a new index is proposed that has several advantages. In particular, the new index allows to evaluate the agreement between raters for each single case (subject or object), and to obtain also a global measure of the interrater agreement for the whole group of cases evaluated. The possibility of having evaluations of the agreement on the single case is particularly useful, for example, in situations where the rating scale is being tested, and it is necessary to identify any changes to it, or to request the raters for a specific comparison on the single case in which the disagreement occurred. The index is not affected by the possible concentration of ratings on a very small number of levels of the ordinal scale.

Keywords: Educational assessment; Interrater agreement; Kappa index; Ordinal rating scales; Statistical dispersion.

1. Introduzione

Classificare soggetti o oggetti in classi (o categorie) predefinite è un'attività piuttosto comune in vari ambiti applicativi ed in particolare in quello

educativo. Basti pensare a tutte quelle situazioni nelle quali è necessario valutare su una scala ordinale le prove di apprendimento linguistico di un gruppo di studenti (ad es. Nuzzo & Bove, 2020) oppure le strategie didattiche messe in atto da un gruppo di docenti (ad es. Graham *et al.*, 2012).

In queste situazioni la scala di valutazione è ritenuta tanto più efficace quanto più il risultato della valutazione dipende soltanto dai valutati e non da coloro che hanno valutato (i valutatori). I valutatori devono risultare quindi interscambiabili e le loro valutazioni pressoché le stesse. Ne consegue che, se consideriamo il caso particolare di due soli valutatori ciascuno dei quali valuta tutti i soggetti (o oggetti) da valutare, ed indichiamo con x ed y le rispettive valutazioni, interesserà analizzare in quale misura le due valutazioni soddisfano la relazione $x = y$ (accordo assoluto). La scala di valutazione sarà efficace se caratterizzata da elevato accordo assoluto tra i valutatori.

Il concetto di accordo assoluto non deve essere confuso con quello di consistenza, che riguarda invece la concordanza delle valutazioni ed è utile quando si deve determinare una graduatoria dei soggetti valutati (ad esempio, la graduatoria dei candidati per una borsa di dottorato o per una posizione lavorativa). Due valutazioni x ed y che soddisfano, ad esempio, la relazione $y = a + bx$ (con a e b non nulli) sono consistenti e perfettamente correlate ma caratterizzate da un basso livello di accordo assoluto. In termini più ampi, possiamo anche dire che il concetto di accordo assoluto non deve essere confuso con il concetto di associazione statistica, quest'ultimo è più generale ed è connesso sia al concetto di accordo che a quello di disaccordo (se c'è un perfetto accordo assoluto c'è anche massima associazione ma non vale il viceversa).

Per analizzare l'accordo tra valutazioni sono state proposte varie strategie (per una rassegna si veda, ad esempio, von Eye & Mun, 2005 o Shoukri, 2011), nel seguito faremo riferimento alla situazione in cui interessi sintetizzare il livello di accordo assoluto con un singolo valore, ottenuto attraverso il calcolo di un opportuno indice. L'attenzione sarà ristretta al caso di valutazioni espresse su scala ordinale nelle quali ciascun valutatore valuta tutti i soggetti (o oggetti) da valutare.

2. Approcci alla costruzione di misure di accordo assoluto per valutazioni su una scala ordinale

Nel seguito trattiamo prima il caso in cui la misura di accordo assoluto riguarda due soli valutatori, e successivamente generalizziamo lo studio al caso di più valutatori.

2.1. *Due valutatori*

Molte delle misure di accordo assoluto utilizzate nel caso di due valutazioni espresse su una scala ordinale costituiscono estensioni dell'indice Kappa (Cohen, 1960), proposto per valutare il livello di accordo assoluto tra due valutazioni espresse su una scala nominale. Richiamiamo le principali caratteristiche dell'indice Kappa con riferimento ai dati della *Tabella 1*, nella quale sono riportati i giudizi di due valutatori sulle prove linguistiche di un gruppo di 40 alunni di una scuola, espressi su una scala ordinale a 4 livelli (ciascun livello è assegnato sulla base di una descrizione delle caratteristiche da riscontrare nella prova linguistica, la prova è tanto migliore quanto più alto è il livello assegnato).

Tabella 1. – Distribuzione dei giudizi di due valutatori su 40 prove linguistiche.

	Valutatore 2				
Valutatore 1	Livello 1	Livello 2	Livello 3	Livello 4	Totale
Livello 1	7	1	1	0	9
Livello 2	2	2	5	0	9
Livello 3	0	1	12	0	13
Livello 4	0	0	6	3	9
Totale	9	4	24	3	40

L'indice Kappa, essendo stato proposto per scale nominali, assume una definizione di accordo di tipo dicotomico (presenza o assenza di accordo), secondo la quale l'accordo o c'è (i 24 giudizi conteggiati lungo la diagonale principale della tabella), o non c'è (tutti gli altri giudizi extra-diagonali). Secondo tale assunzione, la probabilità (che indichiamo con P_a) di riscontrare accordo tra i giudizi dei due valutatori su una prova linguistica scelta a caso sarà rappresentata dalla proporzione dei casi conteggiati lungo la diagonale principale della tabella (ossia: 24/40 = 0,6). L'indice Kappa confronta la probabilità di accordo così ottenuta con quella che si sarebbe ottenuta nel caso i due valutatori avessero espresso casualmente in modo indipendente i loro giudizi sulle prove, secondo le rispettive proporzioni marginali della tabella (cioè la probabilità di accordo calcolata sulla tabella di indipendenza tra i giudizi dei due valutatori). In dettaglio, se indichiamo con P_e la probabilità di accordo ottenuta per effetto delle due scelte casuali, l'indice si può esprimere nel seguente modo:

$$\text{Kappa} = \frac{P_a - P_e}{1 - P_e}$$

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-7916-976-9

Kappa assume valore massimo e pari ad 1 nel caso di perfetto accordo tra i giudizi dei due valutatori (infatti in questo caso è $P_a = 1$), valore pari a 0 quando l'accordo riscontrato nella tabella osservata è uguale a quello ottenuto per effetto del caso ($P_a = P_e$) e valori negativi nel caso l'accordo riscontrato nella tabella è minore di quello ottenuto per effetto del caso ($P_a < P_e$). Sebbene siano state fornite indicazioni non sempre coincidenti sulle modalità di interpretazione dei valori assumibili dall'indice, possiamo dire che generalmente valori dell'indice inferiori a 0,6 si riscontrano in corrispondenza a livelli bassi o moderati di accordo, valori compresi tra 0,6 e 0,8 indicano un buon livello di accordo, valori superiori a 0,8 testimoniano un ottimo livello di accordo. Nel caso della *Tabella 1* si ottiene Kappa = 0,44 che indicherebbe un basso livello di accordo qualora la scala di valutazione fosse trattata come nominale.

La scala di valutazione dei due valutatori considerati nella *Tabella 1* tuttavia non è nominale ma ordinale, di conseguenza, ad esempio, il livello di disaccordo tra il giudizio di livello 4 ed il giudizio di livello 3 non può considerarsi analogo a quello tra il giudizio di livello 4 ed il giudizio di livello 1. Per questo motivo sono state proposte varie modifiche all'indice Kappa in modo da poter introdurre una graduazione del livello di accordo riscontrabile nel confronto tra i livelli della scala. A tale scopo, l'indice Kappa *pesato* (Kappa_w o k_w) proposto da Cohen (1968) ed altri indici a questo collegati (si veda ad esempio Warrens, 2012), nel calcolo delle probabilità P_a e P_e tengono conto di tutte le celle della tabella (non solo di quelle della diagonale principale), introducendo un opportuno sistema di pesi che penalizzi le celle più lontane dalla diagonale principale (corrispondenti a livelli della scala più diversificati) rispetto a quelle ad essa adiacenti. La scelta dei pesi viene operata in modo differente a seconda del contesto applicativo e della sensibilità del ricercatore (una panoramica ampia è fornita, ad esempio, in Gwet, 2014, paragrafo 3.5).

Il calcolo dell'indice Kappa_w è effettuato con la stessa formula vista in precedenza per Kappa, nella quale tuttavia le probabilità di accordo dipenderanno dall'insieme di tutte le celle della tabella e dal sistema di pesi prescelto. Le considerazioni sui valori assumibili da Kappa_w e sulle modalità per la loro interpretazione sono analoghe a quelle svolte in precedenza per l'indice Kappa.

Il calcolo dell'accordo tra i due valutatori della *Tabella 1* può quindi essere effettuato in modo più congruo attraverso la versione pesata dell'indice Kappa. Scegliendo uno dei sistemi di pesi maggiormente utilizzato nelle applicazioni (pesi quadratici), si ottiene $\text{Kappa}_w = 0{,}76$. Il livello di accordo tra i due valutatori può quindi essere considerato buono alla luce del valore ottenuto.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-7916-976-9

Non è infrequente nelle applicazioni che il numero dei valutatori sia superiore a due, ed è quindi necessario avere a disposizione delle misure che consentano di trattare più di due valutatori.

2.2. *Più valutatori*

Nel caso di più di due valutatori la costruzione di una misura di accordo può avvenire in diversi modi.

L'approccio maggiormente studiato ed utilizzato si basa ancora sul concetto di accordo tra valutatori e sul calcolo delle rispettive probabilità di accordo. Tuttavia, con più di due valutatori l'analisi dell'accordo tra le valutazioni può essere fatta in vari modi. Ad esempio, se si hanno 3 valutatori possiamo analizzare l'accordo per coppie di valutatori e poi farne una sintesi, oppure analizzare l'accordo contemporaneamente su tutti e tre i valutatori, considerando l'accordo massimo solo quando le tre valutazioni coincidono. Analogamente, nel caso di 4 valutatori l'accordo si può analizzare in corrispondenza alle coppie di valutatori, oppure alle terne o su tutti e quattro contemporaneamente. Quanto detto si può estendere al caso di un maggior numero di valutatori (si vedano, ad esempio, Conger, 1980 e Warrens, 2012).

Nel caso, molto utilizzato nella pratica, in cui si scelga di considerare l'accordo in corrispondenza alle coppie di valutatori, è possibile definire l'indice globale di accordo tra tutti i valutatori come la media aritmetica dei valori dell'indice di accordo rilevato in corrispondenza a tutte le coppie di valutatori (ad esempio, nel caso di tre valutatori, la media dei tre valori di Kappa_w corrispondenti alle tre coppie possibili di valutatori). È da tener presente che per un numero di valutatori maggiore di 4 o 5, il numero degli indici da calcolare in corrispondenza alle diverse possibili coppie di valutatori cresce molto.

Il calcolo degli indici Kappa e Kappa_w si può effettuare con i più diffusi software statistici o nelle librerie denominate Psych e IRR scritte nel linguaggio R e disponibili all'indirizzo https://cran.r-project.org.

Nel caso in cui invece si scelga di analizzare l'accordo su terne di valutatori o quaterne, ecc., la misura di accordo sarà definita sulla base delle probabilità di accordo P_a e P_e calcolate sulla distribuzione di frequenza a tre, quattro o più dimensioni relativa ai giudizi dei corrispondenti valutatori (per un esempio applicativo si veda Warrens, 2012). È da notare che in questi casi la definizione dei pesi da assegnare ai confronti tra terne, quaterne, ecc. di livelli della scala risulta piuttosto complessa ed il software necessario per il calcolo degli indici non è facilmente accessibile.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-7916-976-9

Gli indici finora menzionati sono caratterizzati da alcune limitazioni non trascurabili. Ad esempio, gli indici Kappa e Kappa$_w$ (o gli indici su questi basati) dipendono fortemente dai totali marginali delle tabelle su cui sono calcolati. Per illustrare questo aspetto, riprendiamo l'esempio dei due valutatori considerato in precedenza, ed introduciamo una variazione nella *Tabella 1* riguardante le frequenze lungo la diagonale principale, concentrando i 24 giudizi coincidenti dei due valutatori quasi tutti sul livello 3 della scala, mantenendo inalterate le frequenze delle celle extradiagonali.

Nella nuova *Tabella 2* così costruita, i totali delle righe e delle colonne sono molto più concentrati sul livello 3 rispetto alla *Tabella 1*. Dal punto di vista dell'accordo tra i due valutatori non è cambiato nulla rispetto alla *Tabella 1*, tuttavia il nuovo valore dell'indice calcolato sulla *Tabella 2* è Kappa$_w$ = 0,41, che evidenzia basso accordo tra i due valutatori.

Tabella 2. – Distribuzione dei giudizi della Tabella 1 con variazione sulla diagonale principale.

	Valutatore 2				
Valutatore 1	Livello 1	Livello 2	Livello 3	Livello 4	Totale
Livello 1	0	1	1	0	2
Livello 2	2	0	5	0	7
Livello 3	0	1	23	0	24
Livello 4	0	0	6	1	7
Totale	2	2	35	1	40

Questo piccolo esempio evidenzia la difficoltà dell'indice a rilevare l'accordo tra valutatori quando questo si esprime in larga prevalenza su un numero di livelli della scala di valutazione particolarmente ristretto.

Un altro limite importante degli indici considerati in precedenza è che dipendono dalla definizione dei pesi da assegnare ai confronti tra i livelli della scala ordinale utilizzata, e tale scelta può variare in relazione alla diversa sensibilità del ricercatore. Questo può rendere difficile il confronto tra i risultati ottenuti in diverse applicazioni.

Infine, vale la pena evidenziare che in certe applicazioni, oltre a conoscere il livello di accordo globale tra i valutatori è utile poter individuare in relazione a quali soggetti (o oggetti) valutati si manifesta l'eventuale disaccordo (ad esempio, può interessare sapere rispetto a quali prove linguistiche i giudizi dei diversi valutatori divergono maggiormente). Gli indici precedentemente menzionati non consentono di ottenere misure di accordo riferite al singolo caso analizzato.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-7916-976-9

Al fine di superare le limitazioni precedentemente evidenziate, nel prossimo paragrafo vengono richiamate le principali caratteristiche di un indice recentemente proposto per la misura dell'accordo assoluto tra più valutazioni espresse su scala ordinale.

3. Una nuova misura di accordo assoluto per valutazioni su scala ordinale

Per introdurre la nuova misura partiamo da un esempio che riguarda i giudizi di 7 valutatori su una prova linguistica di uno studente, espressi utilizzando la precedente scala a 4 livelli. I giudizi sono riportati nella *Tabella 3*.

Tabella 3. – Giudizi di 7 valutatori su una prova linguistica di uno studente (scala ordinale a 4 livelli).

Codice studente	Valut. 1	Valut. 2	Valut. 3	Valut. 4	Valut. 5	Valut. 6	Valut. 7
1	4	2	4	2	3	4	3

Nota: Valut. = Valutatore.

Piuttosto che lavorare come fatto finora sulla probabilità di accordo tra coppie di valutatori (come fanno, ad esempio, O'Connell & Dobson, 1984), rappresentiamo innanzitutto i dati della *Tabella 3* nella seguente distribuzione di frequenza dei sette valutatori rispetto ai livelli della scala ordinale (*Tab. 4*), e lavoriamo su una misura della tendenza ad assumere livelli diversi in essa riscontrata.

Tabella 4. – Distribuzione dei 7 valutatori rispetto ai livelli della scala.

Livelli della scala	N. valutatori
Livello 2	2
Livello 3	2
Livello 4	3
Totale	7

Una misura generale della tendenza a variare di una variabile qualitativa ordinale (o mutabile ordinale) è presentata in Leti (1983), ed è applicata generalmente alle colonne (le variabili) di una matrice di dati. La novità della proposta che si presenta in questo lavoro, è l'idea di applicare la misura alle righe della matrice dei dati che ha in riga i soggetti (o oggetti) valutati ed in colonna i valutatori. Quindi, la distribuzione di frequenza che si analizza (ad esempio quella riportata nella *Tab. 4*) non è quella dei

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-7916-976-9

casi osservati rispetto ai livelli della scala forniti dal singolo valutatore, ma quella dei valutatori rispetto ai livelli della scala assegnati dall'insieme di tutti i valutatori al singolo soggetto (o oggetto) valutato.

La misura da definire sulla singola osservazione (ad esempio i sette livelli assegnati dai valutatori allo studente della *Tab. 3*) dovrà essere non negativa, assumere valore uno nel caso di assenza di variazione dei livelli (ossia nel caso di accordo assoluto massimo, nel quale i giudizi dei valutatori coincidono tutti), ed assumere valore zero nel caso di disaccordo massimo, nel quale i giudizi si equidistribuiscono soltanto sui due livelli estremi della scala. Riportiamo di seguito la sua espressione generale, con riferimento a un numero N di valutatori ed a K livelli della scala ordinale:

$$\delta = 1 - \frac{2\sum_{j=1}^{K-1} F_j(1 - F_j)}{D_{max}}$$

nella quale F_j è la proporzione cumulata associata al livello j della scala nella distribuzione dei valutatori, e D_{max} è il valore $(\frac{K-1}{2})$ nel caso il numero dei valutatori N sia pari, e $(\frac{K-1}{2}) \times (1 - \frac{1}{N^2})$ nel caso il numero dei valutatori N sia dispari. Nella pratica, per un numero di valutatori N maggiore di una decina le due quantità sono pressoché coincidenti entrambe col valore $(\frac{K-1}{2})$.

I valori assumibili dall'indice δ possono interpretarsi con criterio analogo a Kappa, ossia valori dell'indice inferiori a 0,6 si riscontrano in corrispondenza a livelli bassi o moderati di accordo, valori compresi tra 0,6 e 0,8 indicano un buon livello di accordo, valori superiori a 0,8 testimoniano un ottimo livello di accordo. Per la *Tabella 4*, il valore assunto dall'indice è $\delta = 0{,}39$ che testimonia quindi un basso livello di accordo assoluto tra i sette valutatori.

Nelle applicazioni la misura δ consente quindi di individuare tutte le particolari osservazioni in corrispondenza delle quali l'accordo tra i valutatori è basso. Inoltre, una misura globale di accordo dei giudizi dei valutatori sull'intero gruppo valutato (indicata con δ_{medio}) sarà facilmente ottenibile come media aritmetica dei valori dell'indice δ ottenuti per ciascun singolo caso valutato.

La misura proposta consente di ovviare a molti dei problemi posti dall'utilizzo dell'indice Kappa$_w$ e degli altri indici precedentemente descritti. Ad esempio, se si calcola il valore di δ_{medio} per le *Tabelle 1* e *2*, il risultato è lo stesso ed è pari a $\delta_{medio} = 0{,}81$ che evidenzia un livello di accordo molto buono tra i valutatori in entrambe le tabelle, mentre come abbiamo visto il valore di Kappa$_w$ per la seconda tabella è molto diminuito, rilevando erroneamente un più basso livello di accordo rispetto alla prima tabella.

Per un approfondimento degli aspetti statistici ed applicativi legati all'indice δ proposto, si rinvia ai lavori Bove *et al.* (2018, 2020). In particolare, in Bove *et al.* (2020) si dimostra come, nel caso in cui siano stati

osservati un campione di valutatori ed un campione di soggetti (o oggetti) da valutare, sotto determinate condizioni sia possibile costruire intervalli di confidenza per la stima del valore di δ per le popolazioni da cui sono estratti i campioni, senza ricorre a tecniche *computer intensive* di ricampionamento.

Per il calcolo dell'indice δ è stato predisposto un programma di calcolo nel linguaggio di programmazione R che può essere richiesto agli autori di questo contributo. Il programma legge in input una matrice di dati del tipo soggetti × valutatori e restituisce come output i valori individuali e medio dell'indice δ.

Osserviamo infine che, nel caso di valutazioni espresse su scala quantitativa, sono state proposte diverse misure seguendo approcci simili a quello qui presentato (si confronti ad esempio, LeBreton & Senter, 2008, oppure O'Neill, 2017), mentre i coefficienti di correlazione intraclasse (spesso utilizzati anche per scale ordinali tipo Likert) presentano alcuni degli svantaggi evidenziati in precedenza per l'indice Kappa_w.

Riferimenti bibliografici

Bove, G., Nuzzo, E., & Serafini, A. (2018). Measurement of interrater agreement for the assessment of language proficiency. In S. Capecchi, F. Di Iorio, & R. Simone (Eds.), *ASMOD 2018. Proceedings of the Advanced Statistical Modelling for Ordinal Data Conference* (pp. 61-68). Napoli: Università di Napoli Federico II, FedOAPress.

Bove, G., Conti, P. L., & Marella, D. (2020). A measure of interrater absolute agreement for ordinal categorical data. *Statistical Methods & Applications*. https://doi.org/10.1007/s10260-020-00551-5

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 213-220.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213-220.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*, 322-328.

Graham, M., Milanowsky, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Washington, DC: Center for Educator Compensation Reform (CECR), US Department of Education.

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics, LLC.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815-852.

Leti, G. (1983). *Statistica descrittiva*. Bologna: il Mulino.

Nuzzo, E., & Bove, G. (2020). Assessing functional adequacy across tasks: A comparison of learners' and native speakers' written texts. *EuroAmerican Journal of Applied Linguistics and Languages*, *7*(2), 9-27.

O'Connell, D. L., & Dobson, A. J. (1984). General observer-agreement measures on individual subjects and groups of subjects. *Biometrics*, *40*(4), 973-983.

O'Neill, T. A. (2017). An overview of interrater agreement on Likert scales for researchers and practitioners. *Frontiers in Psychology*, *8*, 777. https://doi.org/10.3389/fpsyg.2017.00777

Shoukri, M. M. (2011). *Measures of interobserver agreement and reliability*. Boca Raton, FL: Taylor and Francis Group.

von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement: Manifest variable methods.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Warrens, M. J. (2012). Equivalences of weighted kappas for multiple raters. *Statistical Methodology*, *9*, 407-422.

Riassunto

Vari metodi per analizzare l'accordo tra valutatori sono stati proposti ed applicati in contesti di ricerca che hanno riguardato l'educazione, la psicologia, la sociologia e le discipline mediche. Dopo una sintetica panoramica delle misure di accordo assoluto maggiormente utilizzate per valutazioni espresse su scala ordinale, in questo lavoro si presenta un nuovo indice che ha diversi vantaggi rispetto a quelli esistenti. In particolare, consente di valutare il livello di accordo tra le valutazioni riferite a ciascun singolo soggetto (o oggetto), oltre che a costruire una misura globale dell'accordo assoluto dei valutatori su tutto il gruppo valutato. La possibilità di disporre di valutazioni dell'accordo sul singolo caso risulta particolarmente utile, ad esempio, nelle situazioni in cui la scala di valutazione è in fase di sperimentazione, ed è necessario individuare eventuali sue modifiche, oppure per richiedere ai valutatori un confronto specifico sul singolo caso nel quale si è verificato il disaccordo. L'indice, inoltre, non risente della eventuale concentrazione delle valutazioni su un numero molto ristretto di livelli della scala ordinale utilizzata.

Parole chiave: Accordo tra valutatori; Dispersione statistica; Indice Kappa; Prove di apprendimento; Scale ordinali.

How to cite this Paper: Bove, G., & Marella, D. (2021). Accordo assoluto tra valutazioni espresse su scala ordinale [Interrater absolute agreement for ordinal rating scales]. *Journal of Educational, Cultural and Psychological Studies*, *23*, 239-248. DOI: https://dx.doi.org/10.7358/ecps-2021-023-boma