Ci sono inesattezze nelle standardizzazioni italiane di WISC-IV e WPPSI-IV? (ovvero perché le case editrici dovrebbero essere trasparenti sui dati normativi)

Arjuna Guḍākesha

Author Note

Correspondence concerning this article should be addressed to Arjuna Guḍākesha, Hastinapura, Kuru Kingdom, Bhārat-India.

Abstract

Le scale Wechsler italiane per prescolari e bambini (WISC-IV e WPPSI-IV) potrebbero sovrastimare il QI, in modo variabile a seconda della fascia di età. Questo è quanto suggerito da un'attenta analisi dei contributi italiani alla taratura. Il caso della WISC-IV era già stato sollevato da alcuni autori italiani anni fa, mentre quello della WPSSI-IV viene esposto in questo articolo per la prima volta. Queste elaborazioni evidenziano l'importanza che i dati normativi siano resi disponibili nella loro interezza affinché la comunità scientifica li valuti indipendentemente e, se necessario, sollevi criticità, fornisca raccomandazioni ed eventualmente stimoli dovute correzioni. La politica editoriale implementata da Giunti Psychometrics negli ultimi anni purtroppo non lo renderà più possibile, ed è in contraddizione coi principi dell'*Open Science* crescentemente accettati dalla comunità scientifica internazionale. Questo è specialmente grave considerata la grande rilevanza non solo scientifica, ma anche sociale, delle diagnosi poste usando questi strumenti.

Keywords: scale Wechsler; WISC-IV; WPPSI-IV; Giunti Psychometrics; trasparenza

Word count: 2319

Ci sono inesattezze nelle standardizzazioni italiane di WISC-IV e WPPSI-IV? (ovvero perché le case editrici dovrebbero essere trasparenti sui dati normativi)

Introduzione

Errare è umano, nascondere gli errori è diabolico. È anche per questo che nella ricerca scientifica si sono recentemente imposte le pratiche della cosiddetta "*Open Science*" (es. McKiernan et al., 2016). *Open Science* significa condividere tutta l'informazione: non solo le statistiche descrittive e quelle minime sufficienti a riprodurre i risultati, ma possibilmente tutti i dati e i materiali. Wicherts (2016) ha mostrato che le pubblicazioni scientifiche uscite su riviste che chiedono maggiori standard di trasparenza e *Open Science* contengono effettivamente meno errori. Sarà forse che i ricercatori controllano bene i propri dati, materiali e analisi, se sanno che poi passeranno al vaglio dell'intera comunità scientifica. Esporsi a questo controllo può non piacere al singolo ricercatore, ma tutti riconoscono che sia di grande beneficio per la ricerca scientifica, perché permette di individuare e correggere eventuali errori. C'è poco da fare: la Scienza procede soprattutto per errori e correzioni.

Giunti Psychometrics è una casa editrice privata, non fa ricerca di base e non è una rivista scientifica. Tuttavia raccoglie ed elabora dati, e pubblica strumenti di rilevanza scientifica. Le diagnosi poste con questi strumenti, inoltre, hanno un rilevante impatto sociale e per il benessere degli individui. In forte controtendenza rispetto alla comunità scientifica internazionale, Giunti Psychometrics sta implementando una politica editoriale contraria all'*Open Science*: non solo non rende disponibili i dati raccolti a chi acquista lo strumento, ma non fornisce nemmeno più i dati normativi, le tabelle di conversione grezzo-poderato, le statistiche descrittive, le matrici di correlazione, le statistiche minime per verificare autonomamente la credibilità e attendibilità degli strumenti pubblicati. Ne è un notevole esempio la scala Wechsler per prescolari [WPPSI-IV; Saggino et al. (2019)], e purtroppo ci aspettiamo che segua lo stesso trend anche l'importante scala Wechsler per

bambini (WISC-V, in prossima uscita). Nella WPPSI-IV le uniche statistiche descrittive disponibili riguardano i punteggi ponderati/standardizzati (di limitata utilità in quanto riscalati su una media prefissata), ma ovviamente quelle di interesse riguardano i punteggi grezzi.

In questo articolo viene mostrato, con riferimento alle edizioni italiane di WPPSI-IV (Saggino et al., 2019) e WISC-IV (Orsini et al., 2012), che la disponibilità di alcune informazioni permette, a uno sguardo attento, di rilevare incongruenze e possibili errori. Questa operazione diventerà tanto meno possibile, in futuro, quanto più le informazioni verranno nascoste, e ricercatori, clinici e professionisti saranno costretti ad affidarsi esclusivamente e alla cieca agli algoritmi "black box" dello *scoring online*.

Il caso della WISC-IV

Qualche anno fa, alcuni autori italiani (Giofrè et al., 2017) avevano rilevato apparenti incongruenze tra i trend di sviluppo delle abilità cognitive riportati dalle edizioni UK (ma anche americana) della WISC-IV e quelli riportati dall'edizione italiana. In pratica, mentre le edizioni internazionali suggerivano traiettorie di crescita dei punteggi medi parzialmente curvilinee, quelle italiane risultano lineari per la maggior parte dei subtest. Il confronto, per ovvi motivi, è stato limitato ai subtest non-verbali in cui un confronto diretto, almeno tra popolazioni europee, dovrebbe essere possibile. È importante notare che i punteggi di partenza (a 6 anni) e di arrivo (a 17 anni) sono comunque pressoché identici, e l'anomalia è quindi limitata al trend di sviluppo osservato all'interno dell'intervallo.

Le traiettorie di crescita teoricamente attese nelle abilità cognitive, specialmente per l'intelligenza fluida, dovrebbero essere curvilinee. Ci si attende una più rapida crescita fino all'inizio dell'adolescenza, e poi un proseguimento della crescita, ma progressivamente rallentato fino all'inizio dell'età adulta. Ciò è stato osservato internazionalmente, per esempio, nelle matrici di Raven (cf. Fry & Hale, 2000). È quindi strano che nella taratura

italiana l'andamento dei punteggi medi appaia per lo più lineare. I trend di sviluppo nella WISC-IV italiana sono mostrati sotto in Figura 1.

In base a quanto detto sopra, gli autori italiani avevano concluso che se le deduzioni fossero corrette la WISC-IV sovrastimerebbe i punteggi di QI. Viene notato che "*un individuo [tra gli 11 e i 14 anni] con un punteggio in media in Inghilterra, confrontato con norme italiane avrebbe un indice di ragionamento percettivo compreso tra 111 e 117 (…) e di QI tra 105 e 106 (…)*" (Giofrè et al., 2017, p. 150). Ad ora non si è a conoscenza di repliche da parte di Giunti Psychometrics. Si noti comunque che questa elaborazione è stata possibile solo perché, al tempo, Giunti Psychometrics ancora pubblicava le tabelle di conversione grezzo-ponderato, rendendo possibile tracciare le traiettorie medie di sviluppo. Oggi, senza dati normativi in chiaro, questa elaborazione non si potrebbe più fare.

Il caso della WPPSI-IV

Un'attenta lettura del contributo alla taratura italiana della WPPSI-IV (Saggino et al., 2019)[1] suggerisce che anche qui possa esserci sovrastima dei punteggi, più o meno marcata a seconda delle fasce di età, anche se per ragioni diverse. Nella taratura italiana si parla di un "campione totale (N = 1025)" suddiviso in un "campione normativo (N = 824)" e in un "campione clinico (N = 201)". La Tabella 2-2 della taratura riporta la stratificazione in base all'età. Il campione clinico include ADHD, autismo, disabilità intellettiva, disturbo del linguaggio e DSA. Le percentuali dell'insieme dei casi clinici rispetto al campione totale comunque non sono uniformi per fasce di età, e sono rappresentate sotto in Figura 2. Come si nota, in una fascia di età la percentuale di casi clinici raggiunge addirittura il 30%, mentre in quella subito precedente era solo il 9%.

Naturalmente, come già fatto per la precedente WPPSI-III, ci si aspetta che solo il "campione normativo" (da cui sono stati preliminarmente esclusi tutti i casi con

[1] L'edizione consultata è stata distribuita commercialmente e riporta in calce "2019, Giunti Psychometrics S.r.l. - Firenze"

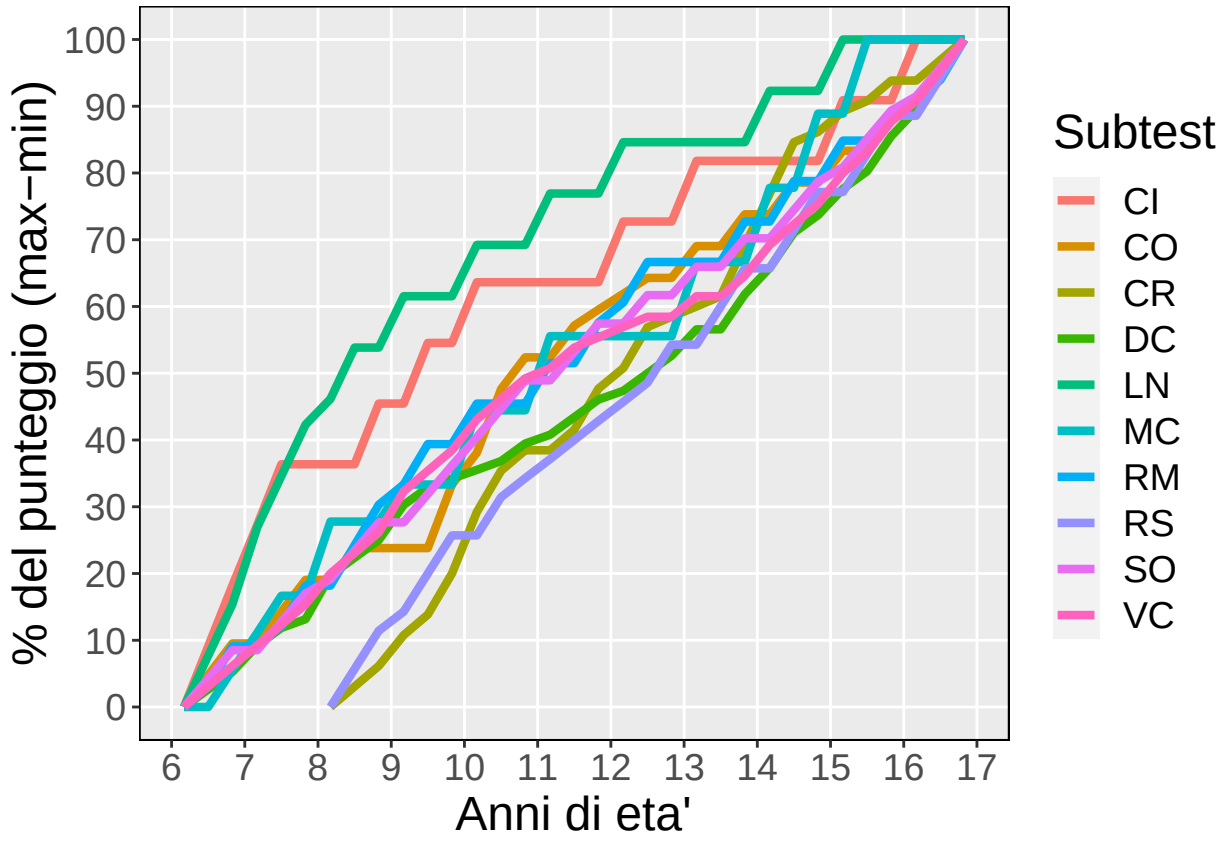

Figure 1

Traiettorie di sviluppo nei 10 subtest fondamentali della WISC-IV, edizione italiana. Il punteggio è riscalato in modo che: 6 anni = 0%, 17 anni = 100%. Si nota che per 8 dei 10 subtest le traiettorie di sviluppo sono decidamente lineari (le uniche due eccezioni sono Riordinamento Lettere-Numeri e Concetti Illustrati.

certificazione) sia stato usato per la derivazione delle norme. Questo è anche quanto sembra evincersi dal testo. Tuttavia, le successive Tabelle 3-1 e 3-4 della taratura contraddicono questa aspettativa. In entrambe vengono riportati i punteggi ponderati/standardizzati "per il campione totale (N = 1025)", quindi inclusivo dei casi clinici. E in entrambe i punteggi sono esattamente attorno al dato normativo, ovvero attorno a 10 per i subtest e attorno a 100 per gli indici. Poiché il "campione totale" include un notevole 20% di casi clinici (ma in alcune fasce di età si rasenta o si raggiunge il 30%), ci saremmo aspettati punteggi mediamente più bassi di così. La Tabella 3-7, che riporta i dati medi dei punteggi ponderati del campione clinico, mostra infatti che questo campione (preso nell'insieme) è mediamente compromesso in tutti i subtest.

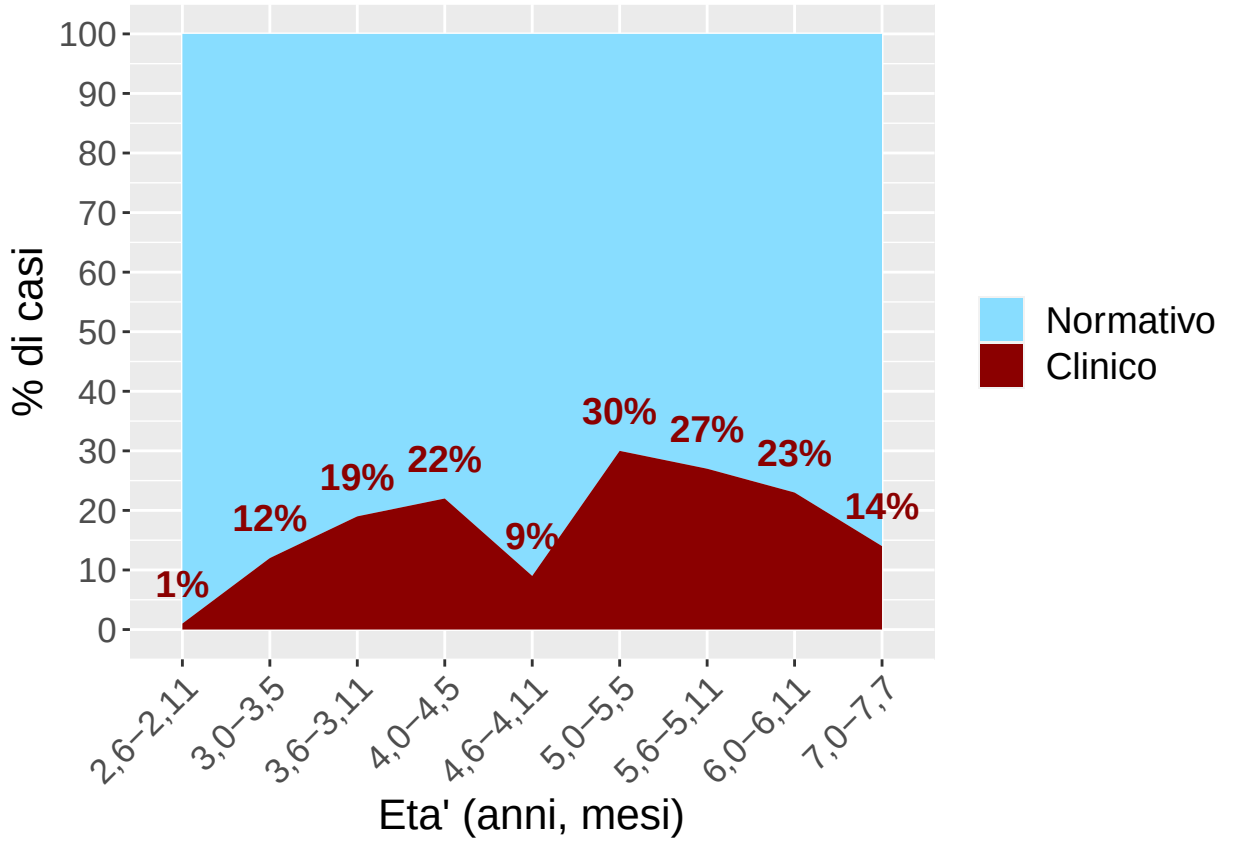

Figure 2
Distribuzione dei casi clinici, divisa per fascia di età, all'interno del 'campione totale' (N = 1025) dell'edizione italiana della WPPSI-IV.

Il campione clinico è chiaramente "sovracampionato", perché raccolto appositamente tramite canali a parte, non identificato casualmente attraverso uno screening della popolazione. Di certo un 30% di casi clinici appositamente aggiunti nella fascia di età "5,0-5,5 anni" appare un forte sovracampionamento, anche perché parte dei restanti casi "normativi" potranno comunque essere in seguito identificati come DSA o con altri disturbi, e semplicemente non avere ancora le condizioni per una diagnosi. È quindi possibile che la sovrarappresentazione dei casi clinici sia ancora superiore a quanto mostrato in Figura 2.

Se le Tabelle 3-1 e 3-4 della taratura mostrano chiaramente che il "campione totale (N = 1025)", ha punteggi ponderati attorno a 10 e standardizzati attorno a 100, ma in esso i casi clinici sono sovracampionati e hanno punteggi inferiori, questo significa che i casi con sviluppo normotipico avranno punteggi sovrastimati. Di certo, se la WPPSI-III NON

include casi clinici nel campione di standardizzazione mentre la WPPSI-IV lo fa, significa che lo stesso bambino valutato con la WPPSI-IV oggi sarà sovrastimato rispetto a quando era stato valutato con la WPPSI-III in precedenza. Significa anche che lo stesso bambino, quando passa da una fascia di età all'altra, potrebbe essere sovrastimato in misura diversa dalla stessa WPSSI-IV, dando luogo a incoerenze psicometriche.

Si può stimare *quanto ampia* sia la sovrastima operata dalla WPPSI-IV? È difficile dirlo, perché non sono disponibili le curve di sviluppo dei punteggi grezzi, né le matrici di varianza e covarianza dei punteggi grezzi e standardizzati proprio a causa delle nuove politiche editoriali di Giunti Psychometrics. Ma possiamo provare a fare delle supposizioni. Prendiamo il QI totale: dovrebbe avere media = 100 e deviazione standard = 15. Da un primo sguardo alla Tabella 3-7 si può dire che i punteggi ponderati ai subtest siano compatibili con un QI attorno a 85. Poiché esso rappresenta il 20% del campione totale, se lo rimuovessimo del tutto ci aspetteremmo che il QI dei restanti casi sia di circa 104. Da questa grossolana approssimazione, potremmo concludere che la WPPSI-IV sovrastimi di circa 4 punti di QI rispetto alla WPPSI-III.

Facciamo ora un conto più raffinato. Per tutte le analisi è stato usato il software R (R Core Team, 2023) (libero e gratuito). Per generare i dati è stato usato il pacchetto "MASS" (Venables & Ripley, 2002). Per ciascuna fascia di età e per il campione totale simuliamo 10^6 osservazioni di casi "clinici" ed altrettante di casi "normativi" (i numeri ampi servono solo a garantire stabilità dei risultati). Per il campione "normativo" generiamo ogni volta punteggi ai subtest distribuiti normalmente con M = 10 e DS = 3, mentre per il campione clinico generiamo punteggi con M (media) e DS (deviazione standard) dei punteggi ponderati come riportati in Tabella 3-7 (combinando tutti i casi assieme ma rispettandone le proporzioni). In tutti i casi simuliamo punteggi correlati con r = 0.50 (come detto sopra, non sappiamo però quale sia la vera correlazione tra i subtest, perché i manuali non la riportano). A questo punto riscaliamo la somma dei ponderati del campione "normativo" in modo che QI totale abbia M = 100 e SD = 15. Usiamo questi

dati normativi per calcolare i QI dei casi "clinici". A questo punto, tenendo conto della percentuale di casi "clinici" in ciascuna fascia di età (cf. Figura 2), calcoliamo per differenza quanto debba essere il QI medio della restante parte di campione "tipico" per garantire di avere sempre una media pari a 100. La differenza tra questo QI medio e il valore 100 è presa come sovrastima per quella fascia di età.

I risultati del calcolo sono come segue. Per il campione complessivo, abbiamo un QI medio dei casi clinici pari a 84.09, una sovrastima per i "tipici" pari a 3.89 punti di QI. La sovrastima attesa per ciascuna fascia di età è mostrata sotto in Figura 3.

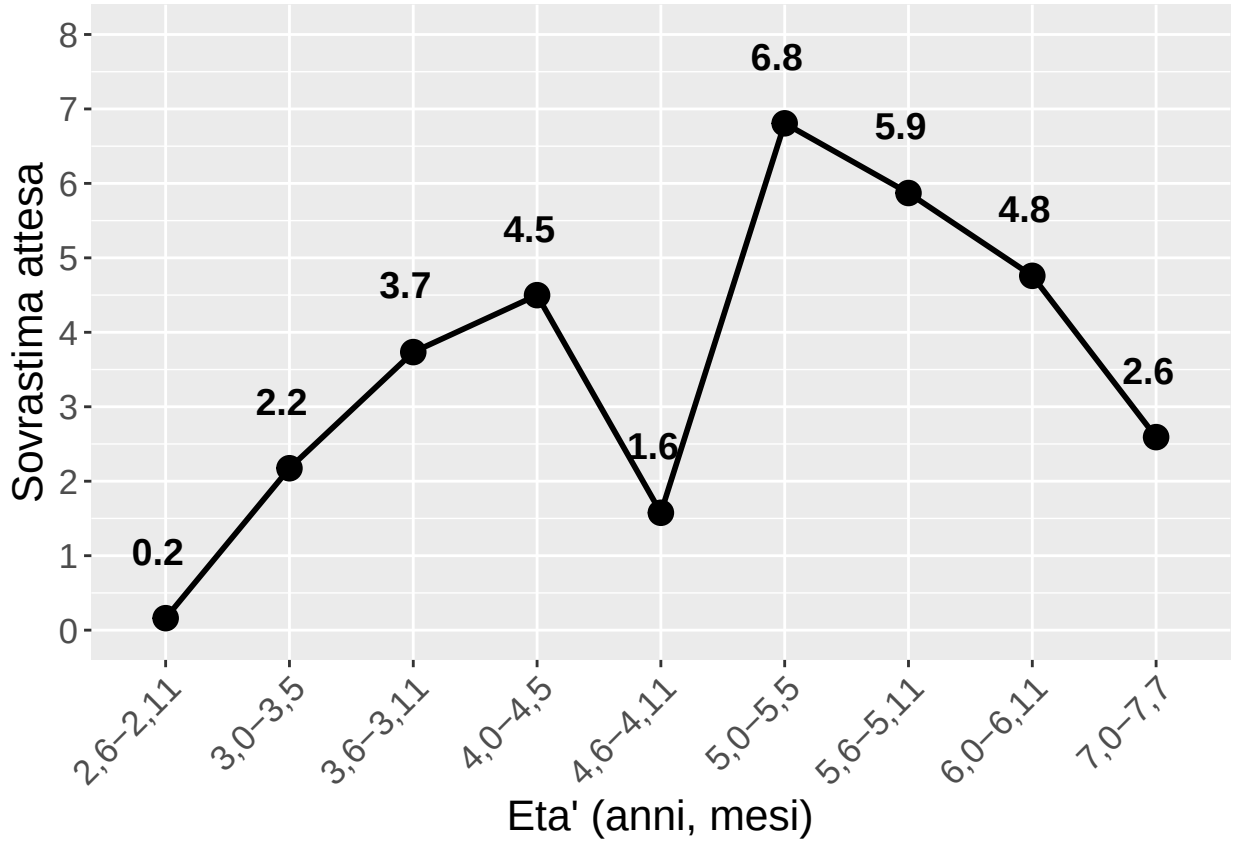

Figure 3
Sovrastima attesa del QI per fascia di età nella WPPSI-IV, edizione italiana.

Sintesi dei risultati

Un'attenta analisi dei contributi alla taratura delle edizioni italiane di WISC-IV e WPPSI-IV suggerisce che entrambe le batterie potrebbero sovrastimare i punteggi

standardizzati. La rilettura di un precedente articolo pubblicato da autori italiani (Giofrè et al., 2017) suggeriva che la WISC-IV sovrastima fino a 5-6 punti di QI in alcune fasce di età, e addirittura un'intera deviazione standard (11-17 punti) per quanto riguarda l'indice di Ragionamento Percettivo in quelle stesse fasce di età. Per ragioni differenti, anche la WPPSI-IV sembra sovrastimare i punteggi standardizzati, con eccessi fino a 5-7 punti di QI in alcune fasce di età.

Nel caso della WPPSI-IV, si potrebbe controbattere che il campione clinico non debba essere escluso dalla standardizzazione in quanto fa parte della popolazione generale. Questo punto ci trova potenzialmente d'accordo. Tuttavia, rimangono alcune incongruenze per quanto riguarda le scale Wechsler italiane. In primo luogo, la procedura di rimuovere preliminarmente i casi certificati in fasce d'età molto precoce è discutibile perché a questa età non tutti i disturbi del neurosviluppo possono essere già stati diagnosticati, e questa criticità varia da fascia a fascia d'età. In secondo luogo, aggiungere casi clinici in grande numero ottenuti attraverso altri canali (diversi da quelli dello screening generale, che garantisce la rappresentatività complessiva) in proporzioni addirittura fino al 30% del campione totale, porta comunque a un sovracampionamento, che è molto problematico per la validità dei dati normativi e porta a sovrastime. Infine, la procedura adottata nella WPPSI-IV appare comunque incongruente con quanto già fatto con la WPPSI-III, portando a possibili incoerenze psicometriche.

Conclusioni

È fondamentale che le case editrici di test psicometrici adottino politiche di trasparenza sui dati normativi e su tutte le statistiche necessarie per controllare validità e credibilità delle standardizzazioni. I dati psicometrici sono di grande rilevanza non solo scientifica, ma anche sociale. È su questi dati che si basano le diagnosi che hanno poi valore legale nonché grande impatto sulla vita e il benessere dei singoli. La segretezza dei dati normativi, in questo contesto, appare molto grave.

Alcune informazioni fondamentali minime sulle standardizzazioni delle batterie in uso (tabelle di conversione grezzo-ponderato, statistiche descrittive *dei punteggi grezzi* divise per fasce di età, matrici di varianza e covarianza divise per fasce di età) dovrebbero sempre essere rese disponibili in modo che la comunità scientifica possa fare valutazioni indipendenti e, se necessario, sollevare criticità, fornire raccomandazioni, e in casi estremi portare a eventuali correzioni. La politica editoriale seguita da Giunti Psychometrics negli ultimi anni, ma purtroppo anche da altre case editrici per alcuni test, è invece in contraddizione con quanto appena detto, e in controtendenza con la ricerca scientifica internazionale che sta adottando come prassi necessarie i principi della trasparenza e dell'*Open Science*.

References

Fry, A. F., & Hale, S. (2000). Relationships among processing speed, working memory, and fluid intelligence in children. *Biological Psychology*, *54*(1-3), 1–34.

Giofrè, D., Toffalini, E., & Provazza, S. (2017). La WISC-IV sovrastima le competenze dei ragazzi italiani? Discrepanze tra la standardizzazione UK e quella italiana della scala. *Psicologia Clinica Dello Sviluppo*, *21*(1), 143–154.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., et al. (2016). How open science helps researchers succeed. *Elife*, *5*, e16800.

Orsini, A., Pezzuti, L., & Picone, L. (2012). *WISC-IV: Contributo alla taratura italiana. [WISC-IV italian edition]*. Florence, Italy: Giunti OS.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Saggino, A., Giacomo, S., & Claudio, V. (2019). *WPPSI–IV wechsler preschool and primary scale of intelligence fourth edition. Contributo alla taratura italiana. [WPPSI-IV italian edition]*. Florence, Italy: Giunti Psychometrics.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Wicherts, J. M. (2016). Peer review quality and transparency of the peer-review process in open access and subscription journals. *PloS One*, *11*(1), e0147913.