



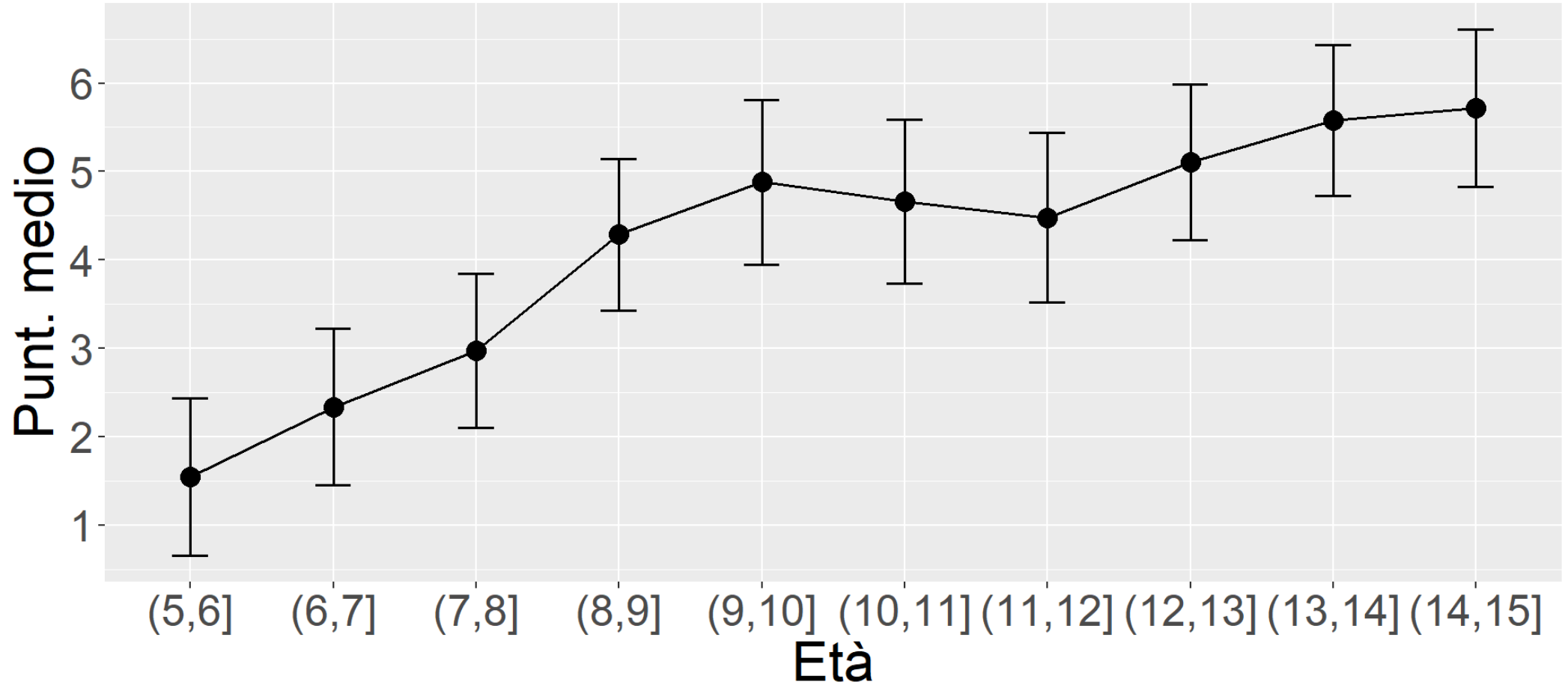
Quando Vorremmo più Dati: Alcuni Esempi

ovvero, se dubitiamo del testing quando è trasparente,
figuriamoci quando non lo è

Enrico Toffalini
DPG - Università di Padova

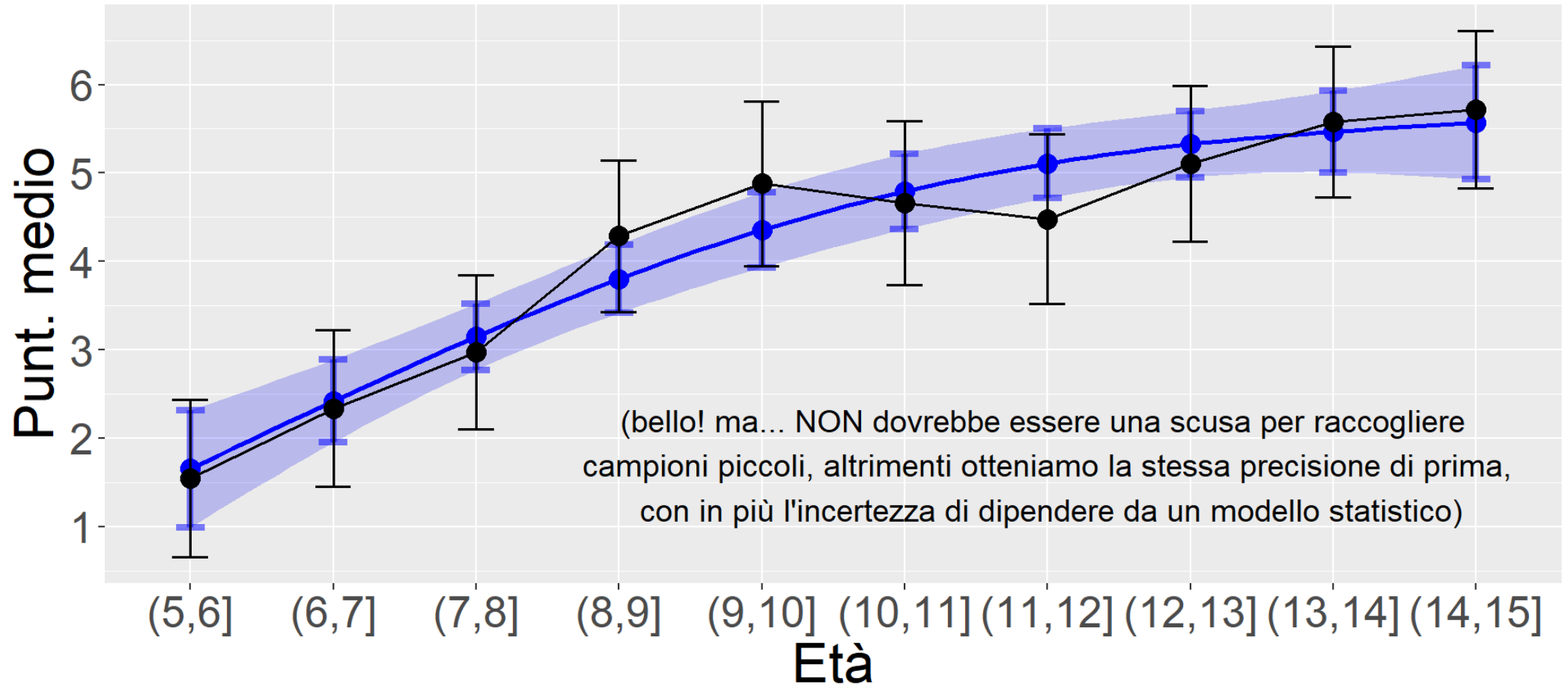
Premessa: W il *continuous norming*

Statistiche descrittive



Premessa: W il *continuous norming*

Statistiche descrittive e Continuous norming



Premessa: *W lo scoring online*

[Styck e Walsh \(2016\)](#), meta-analisi di 27 studi: in media il 99.7% dei protocolli Wechsler contiene almeno un errore

Premessa: *W lo scoring online ?*

[Styck e Walsh \(2016\)](#), meta-analisi di 27 studi: in media il 99.7% dei protocolli Wechsler contiene almeno un errore

MA... scende a 41.2% escludendo mera omissione trascrizione

MA... i professionisti hanno meno rischio di farne almeno uno (34% vs 70% esclusa omissione trascrizione)

MA... protocolli ri-corretti cambiano **mediamente punto di QI**

MA... varianza molto maggiore (SE punti QI) sembra dipendere da **effetto/bias somministratore** ([McDermott et al., 2014](#)), **non risolvibile con scoring online**


Quali casi esempio vediamo ora...

Cosa sappiamo e cosa non sappiamo delle caratteristiche psicometriche di alcuni strumenti di ampio uso:

- **Conners 3** (problemi comportamentali, ADHD)
- **Scale Wechsler: WISC-IV, WPPSI-IV, WISC-V**
(casistica di ogni tipo nello sviluppo)
- **MOXO d-CPT (ADHD)**

Disclaimer: l'elenco NON è esaustivo degli strumenti con possibili problematicità; l'elenco NON implica che le alternative siano necessariamente migliori; la selezione è basata primariamente su esperienza personale; si invitano caldamente gli interessati a verificare autonomamente le informazioni riportate

Conners 3: *scoring online* corregge errori?

 CONNERS 3®	CODICE CLIENTE XXXXXXXXXXXX
	CODICE QUESTIONARIO XXXXXX
	DATA DELLA PROVA <u>2024-01-</u> XX

Dati del bambino

NOME E COGNOME/ID	XXXXXXXXXXXX
GENERE	Maschile
DATA DI NASCITA	<u>XX-12-2009</u>
ETÀ	<u>11</u>
ORDINE DI GENITURA	
CLASSE FREQUENTATA	<u>1° anno università</u>

test effettuato a gennaio 2024 su psy-portal.giuntitesting.com

Conners 3: psicometria dei *punti T* ?

7. Indice Globale Conners 3⁶

Dalla descrizione dell'insegnante, il bambino/adolescente ha ottenuto un punteggio T pari a 124 e un punteggio percentile pari a 99.



Provocazione/Aggressività

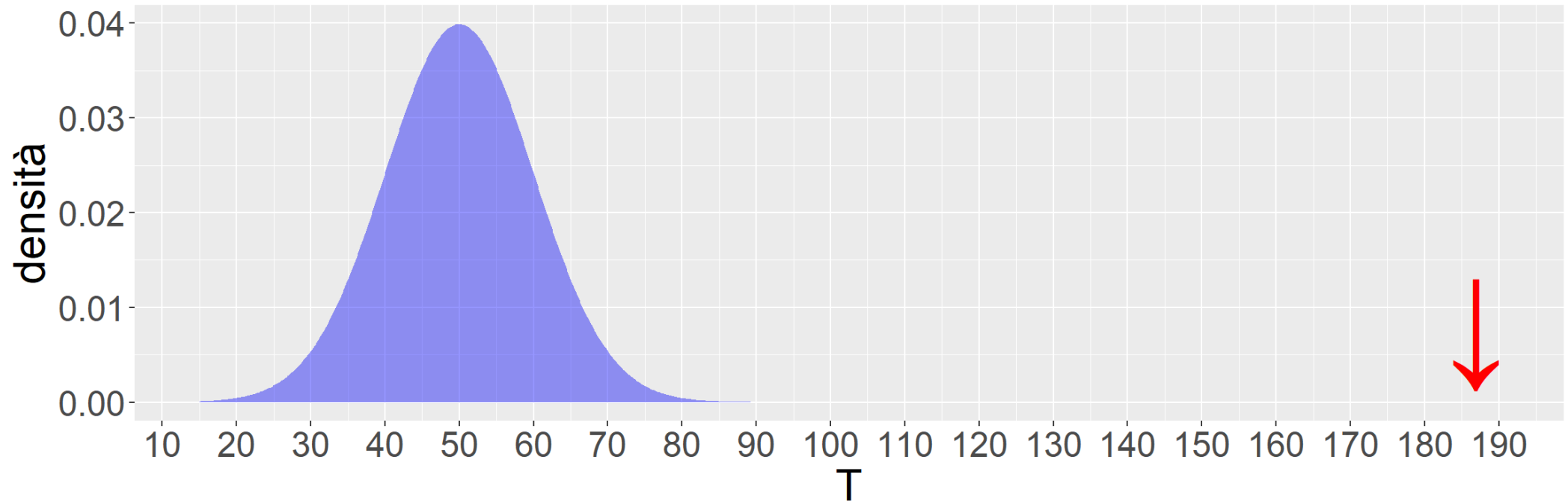


selezione da protocolli di casi reali segnalati a marzo 2024

T = 187 corrisponde a $z = +13.7$: se facesse riferimento alla curva Gaussiana, dovrebbe esistere meno di *un caso su un milione di miliardi di miliardi di miliardi di miliardi di persone* ($1 / 10^{42}$)

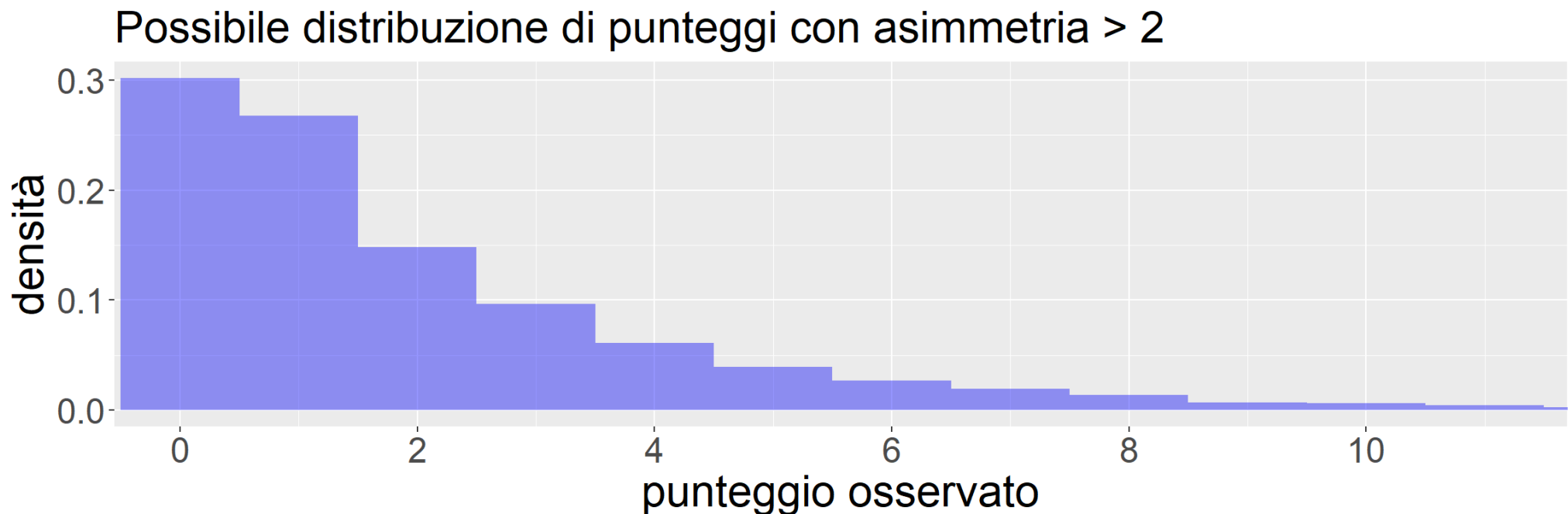
Conners 3: psicometria dei *punti T* ?

Cosa sarebbe $T = 187$ in una distribuzione normale



Conners 3: psicometria dei *punti T* ?

Ricevo un manuale (2017) che mi permette effettivamente di ricalcolare (*esempio virtuoso!*) $T = 187 / z = +13.7...$ basandomi su un sottocampione di $n = 24$ (in base a età e genere). I valori di asimmetria per alcune scale raggiungono e superano 2



Conners 3: psicometria dei *punti T* ?

(perdonate il pippozzo)

La metrica dei punti T e z è interpretabile solo rispetto alla **curva Gaussiana**, dove *Media* e *Deviazione Standard* sono effettivamente parametri che governano la distribuzione

Altrimenti... sì, potete sempre calcolare T e z, ma non vi dicono la collocazione della persona nella popolazione rispetto al costrutto di interesse (es.

“*provocazione / aggressività*”), ma solo del suo punteggio rispetto agli altri punteggi che, date certe distribuzioni estremamente asimmetriche, riflettono in modo forte le caratteristiche della prova (ma a noi interessa la persona)

Conners 3: SINTESI

- *Scoring online non corregge né segnala palesi incongruenze* dovute a errore umano (almeno a gennaio 2024), mancando uno dei suoi principali vantaggi
- Distribuzioni **estremamente asimmetriche**, e tuttavia *scoring* basato su punti T (**punti z**); inoltre, il manuale “valida” numerosi modelli fattoriali EFA e CFA (che assumono normalità) a 4 e 5 fattori usando questi punteggi
- **Dati normativi trasparenti, ma sottocampioni estremamente piccoli** (molti $n = 10 - 40$ osservazioni) su cui si basano dati normativi, senza *continuous norming*

WISC-IV: rischio sovrastima?

Esempio parzialmente virtuoso di trasparenza (fornisce tabelle conversione grezzo-ponderato, ma non statistiche descrittive complete), la **WISC-IV** permette *indirettamente* di valutare **traiettorie di sviluppo**



Psicologia clinica dello sviluppo
Fascicolo 1/2017, aprile

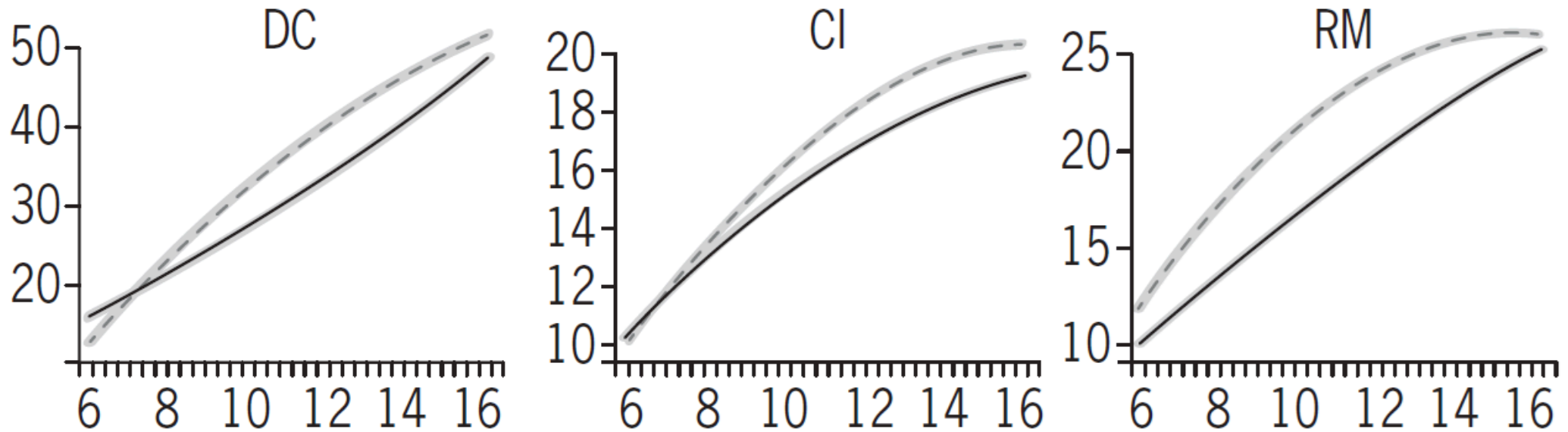
[Indice](#) | [Precedente](#) | [Successivo](#)

David Giofrè,  Enrico Toffalini,  Serena Provazza

La WISC-IV sovrastima le competenze dei ragazzi italiani? Discrepanze tra la standardizzazione UK e quella Italiana della scala

pp: 143-154 | DOI: 10.1449/86189

WISC-IV: rischio sovrastima?

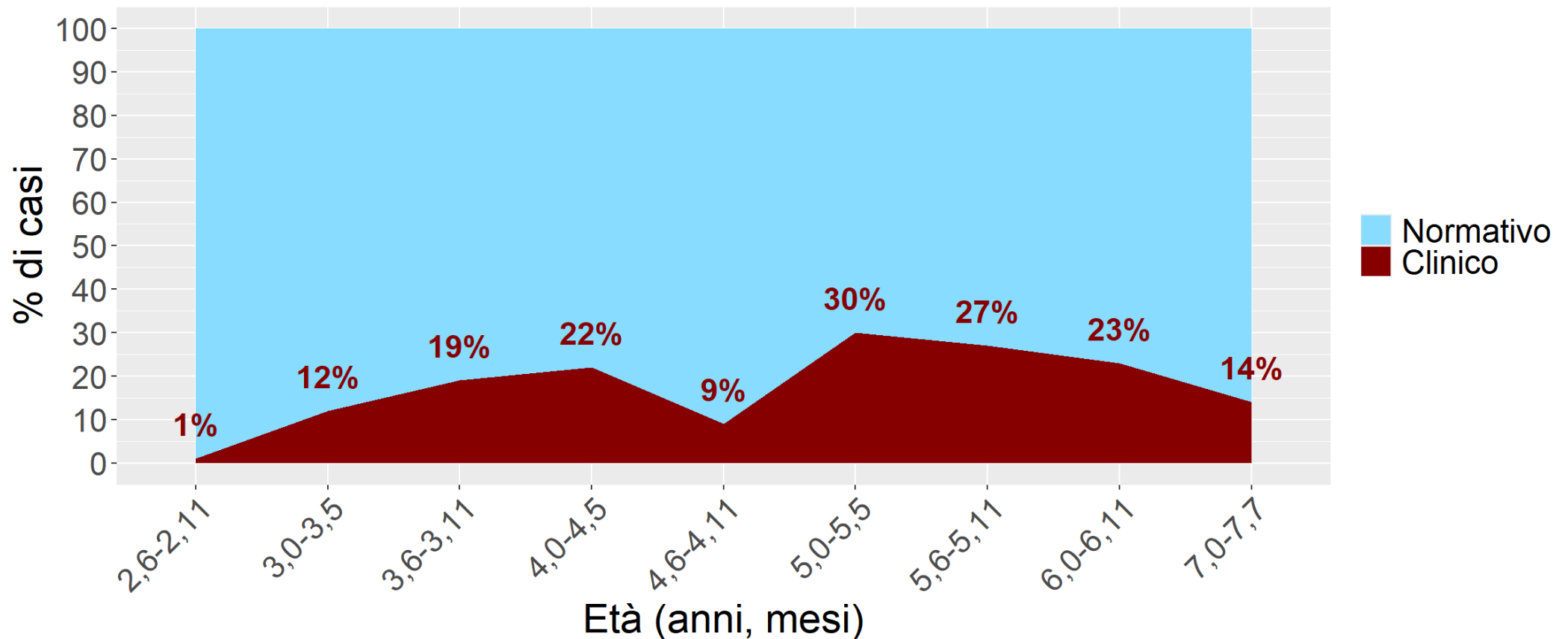


Traiettorie UK curvilinee come ci si aspetta tipicamente e teoricamente (e coerentemente con le norme americane); quelle italiane invece sembrano “stranamente” lineari.

“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)

WPPSI-IV: once again, rischio sovrastima?

La versione per prescolari (WPPSI-IV) potrebbe anch'essa sovrastimare, secondo [un recente report](#), per motivi diversi



WPPSI-IV: once again, rischio sovrastima?

- Il “*campione totale* ($N = 1025$)” include 20% casi clinici (disturbi del neurosviluppo; $n = 201$) i quali in base a Tabella 3-7 potrebbero avere QI medio
- Tabelle 3-1 e 3-4 suggeriscono che il “*campione totale* ($N = 1025$)” abbia punteggi ponderati e standardizzati, facendo sospettare che il dato medio dipenda da una pletora di casi clinici (la WPPSI-III li escludeva esplicitamente)
- Una rielaborazione suggerisce che, se vero, **WPPSI-IV sovrastimerebbe** (almeno rispetto a WPPSI-III) in media **+4 punti di QI**, e fino quasi a **+7** a seconda della fascia di età

WISC-V: ???

Del vero “oggetto della contesa” sappiamo poco: non abbiamo le statistiche descrittive dei dati grezzi (a febbraio 2024 risultavano riportate solo per i ponderati, che sono poco utili), né le matrici di correlazione divise per fascia di età (solo la complessiva), né le tabelle di conversione grezzo-ponderato. Questo rende impossibile indagare traiettorie di sviluppo, sapere se overfittano, valutarle rispetto a traiettorie attese, fare qualsiasi *double check* delle norme, o verificare l’invarianza di misura per età

Scale Wechsler: SINTESI

- *WISC-IV* e *WPPSI-IV* **potrebbero sovrastimare QI**; impressione clinica con *WISC-IV* conferma che subtest percettivi sono “troppo spesso” punti di forza, indipendentemente da condizione clinica
- Dati normativi forniti sono al meglio *parziali* (*WISC-IV*: conversione grezzo-ponderato, no statistiche descrittive base), **al peggio assenti (*WISC-V*)**, costringendo a cieca fiducia nella casa editrice e nel suo *scoring online*
- *WISC-V*: **comunità accademica e professionale non ha modo di effettuare verifiche indipendenti** della qualità psicometrica in base a informazioni pubblicamente disponibili

Monopolio? Non per forza

Se pensiamo a nomi “*Raven*” e “*Wechsler*”: sì, ci sono diritti di copyright esclusivi

... ma se intendiamo “fattore g ” o “teoria CHC” o “modello gerarchico”, ovviamente NO!

IDS-2: Una valida alternativa (?)

IDS-2 - Intelligence and Development Scales – 2nd Edition

Alexander Grob, Priska Hagmann-von Arx



Sintesi

Finalità:	valutazione dell'intelligenza, delle funzioni esecutive e dello sviluppo
Età/target:	5-20.11 anni
Caratteristiche:	test di performance con utilizzo di immagini e materiali
Composizione:	6 aree funzionali, 30 subtest
Somministrazione:	<i>forma:</i> individuale (carta e matita) <i>tempo:</i> Qi Screening: 10'; Qi Generale: 50'; Qi Approfondito: 90'; Funzioni esecutive: 30'; Abilità psicomotorie: 20'; Competenze socio-emozionali: 15'; Competenze scolastiche: 30-60'; Motivazione: 7'
Scoring:	<i>forma:</i> protocollo di registrazione + scoring online
Psicometria:	punteggi QI, punteggi scalari, punteggi età equivalente,

disclaimer: fornisce conversione grezzo-ponderato in chiaro, ma NON tutte le statistiche descrittive

MOXO d-CPT

Per ultimo, Moxo d-CPT (Hogrefe):

- prova computerizzata, *scoring online*; manuale **fornisce statistiche descrittive, ma solo basate su continuous norming, non quelle osservate**, per indici principali (non per sottoindici); clinici riferiscono che **piattaforma di *scoring* NON restituisce punteggi grezzi dei pazienti valutati!**
- fornisce 4 indici ma **non include modelli fattoriali né statistiche sufficienti** per validarli;
- **promette ottima capacità diagnostica per ADHD** (punteggio combinato, **AUC = 0.91**)

MOXO d-CPT: *troppa* capacità diagnostica?

Arrondo, G., Mulraney, M., ... Cortese, S. (2024). Systematic review and meta-analysis: Clinical utility of Continuous Performance Tests for the identification of Attention-Deficit/Hyperactivity Disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 63(2), 154-171.

<https://doi.org/10.1016/j.jaac.2023.03.011>

“MOXO (...) appeared as an outlier compared to the other CPT versions (...) When taken at face-value, our analyses suggest a higher accuracy for the MOXO compared to other CPTs. However, our confidence in this difference is low and might not replicate in future studies (...) the studies on the MOXO had a high risk of bias (...) was likely to have inflated the clinical utility measures.

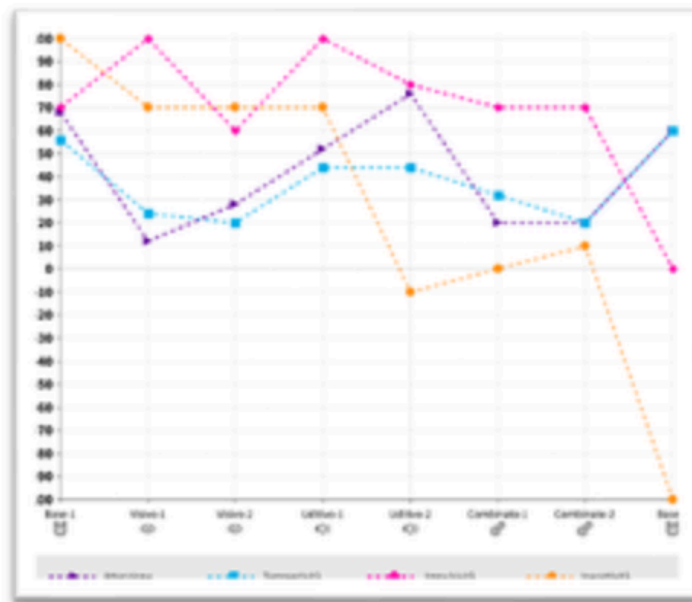
Additionally, authors of MOXO studies were involved in its development and commercialization, suggesting a direct conflict of interest” [NOTA: l'eventuale

conflitto di interesse farebbe riferimento agli autori originali, NON ai curatori italiani]

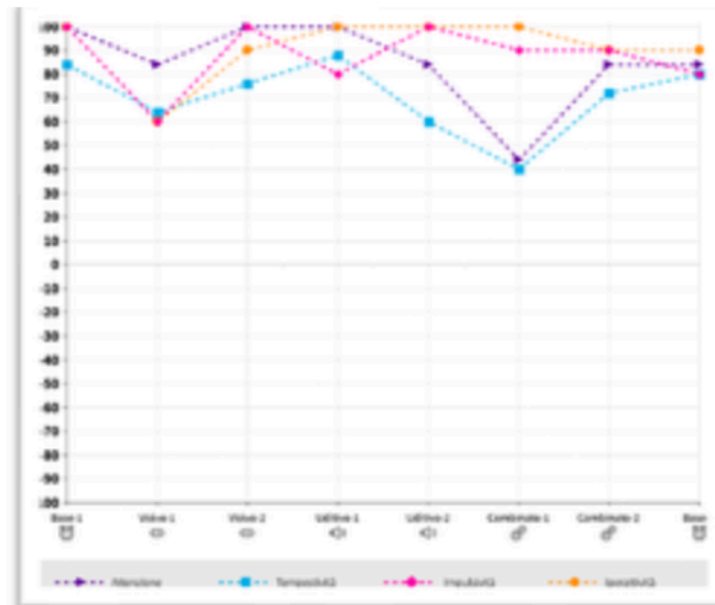
MOXO d-CPT: psicometria dei punti z

z estremi fanno pensare che NON stia misurando il tratto, ma descriva solo il punteggio a una prova

	Attenzione	Tempestività	Impulsività	Iperattività
Pre	$z = -8.09$	$z = -5.36$	$z = -3.17$	$z = -8.95$
Post	$z = -1.76$	$z = -0.56$	$z = -0.15$	$z = 0.16$



Prima somministrazione



Somministrazione post-trattamento

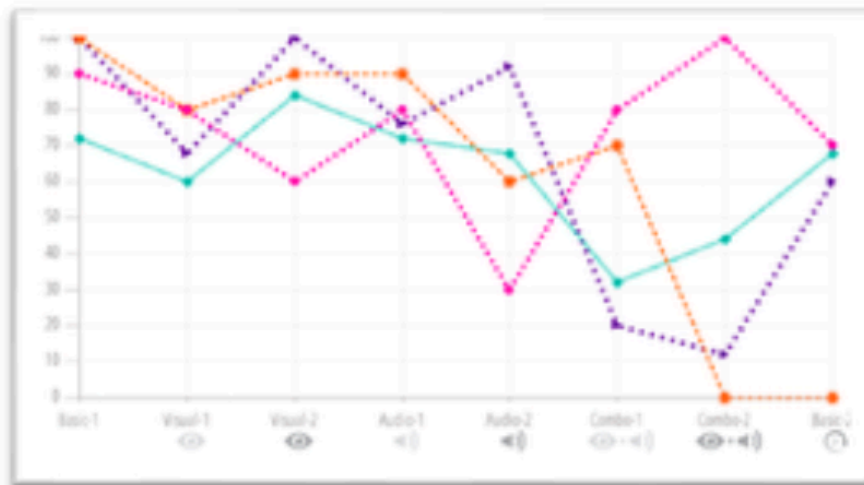
da <https://qi.hogrefe.it/rivista/moxo-d-cpt-profilo-di-funzionamento-e-misura-del-c/>
(verificato ad aprile 2025)

MOXO d-CPT: psicometria del cambiamento

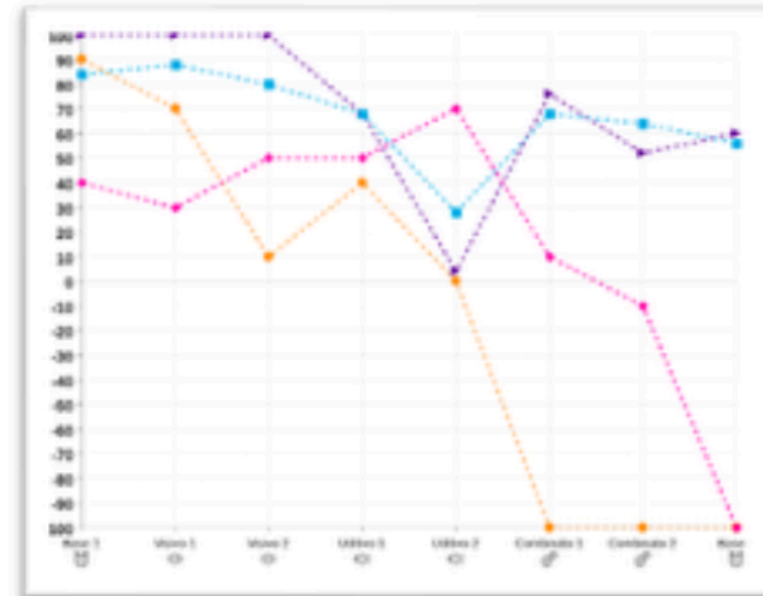
variazioni di molte Dev.St. in un tratto a seguito di intervento si ritengono implausibili

	Attenzione	Tempestività	Impulsività	Iperattività
Pre	$z = -3.98$	$z = -2.4$	$z = -1.73$	$z = -5.40$
Post	$z = -1.42$	$z = 0.21$	$z = -5.20$	$z = 5.78$

+11 pz?



Prima somministrazione



Somministrazione post-trattamento

da <https://qi.hogrefe.it/rivista/moxo-d-cpt-profilo-di-funzionamento-e-misura-del-c/>
(verificato ad aprile 2025)

MOXO d-CPT: SINTESI

- Scoring online con dati parzialmente in chiaro (ma punti z NON possono essere ricalcolati autonomamente; inoltre statistiche descrittive “originali” non disponibili);
- Ottima “capacità diagnostica” messa in dubbio da recente review e meta-analisi;
- Punti z estremi sembrano poco appropriati come indicatore psicometrico, troppo variabili nel tempo e tra indici (il manuale descrive buona stabilità test-retest su sottocampione italiano, ma solo a distanza di 60 minuti);
- *“profilo di funzionamento (...) estremamente dettagliato”* solleva dubbi: il manuale non permette di verificare modello fattoriale, punteggi probabilmente troppo volatili, indici forse poco attendibili

SINTESI delle SINTESI

- *Open Science* → strumenti psicometrici devono riportare non solo conversione grezzo-ponderato, ma anche **tutte le statistiche descrittive sufficienti**, possibilmente dati grezzi anonimizzati, per verifiche indipendenti;
- *Scoring online* e *continuous norming* vanno benissimo ma non esimono da quanto sopra; dovremmo chiedere uno **standard minimo di trasparenza “per legge”** su strumenti *high-stake*;
- In caso contrario, **l'unico garante rimane chi vende gli strumenti** (con eventuali collaboratori), che è (giustamente) **in conflitto di interesse rispetto a evidenza di validità, affidabilità, capacità diagnostica**