Relazione tecnico-scientifica sui chiarimenti forniti dal CISIA nella relazione del 30/11/2023

Ing. Carmelo Pino
Dottore di Ricerca in Ingegneria Informatica

19 dicembre 2023

1 Introduzione

Questa relazione presenta un parere tecnico-scientifico riguardante i chiarimenti forniti dal CISIA su tre questioni specifiche:

- Modalità di calcolo del punteggio equalizzato.
- Numero di quesiti che compongono la banca dati.
- Criteri e modalità di individuazione dei quesiti da sottoporre ai candidati nei diversi turni della stessa sessione e nelle sessioni separate.

La psicometria [1], una disciplina delle scienze comportamentali, fornisce risposte esaustive a tutti e tre i punti in discussione. La psicometria si occupa dello sviluppo, della validazione e dell'analisi degli strumenti di misurazione psicologica, con l'obiettivo principale di quantificare e comprendere tratti psicologici, abilità cognitive e caratteristiche comportamentali attraverso l'impiego di metodi scientifici e approcci statisticamente robusti. Uno degli strumenti più diffusi in psicometria è il test, che può assumere forme diverse, spaziando dalle valutazioni standardizzate alle indagini psicologiche approfondite.

2 Meccanismo di equalizzazione

L'equalizzazione dei test, o *equating*, è una procedura tecnica finalizzata a stabilire punteggi comparabili e significativamente equivalenti tra diverse versioni di un test [2]. Questo aspetto diventa particolarmente rilevante quando si progettano test in forme multiple per diversificare le domande o per adattarle a contesti specifici, garantendo allo stesso tempo coerenza nella valutazione delle competenze e delle caratteristiche misurate. L'equalizzazione è quindi essenziale per interpretare correttamente le misure psicometriche e per effettuare confronti affidabili tra individui, gruppi o nel corso del tempo.

L'esigenza di collegare e bilanciare i risultati dei test deriva dalla necessità di compensare le differenze nella difficoltà delle diverse forme di test. Consideriamo un caso in cui uno studente ottiene un punteggio di 72/100 su una forma di test (A), mentre un altro studente ottiene 74/100 su un'altra forma di test (B). Senza l'equalizzazione, diventa problematico stabilire se il secondo studente sia effettivamente più competente o se la sua forma di test presentava semplicemente domande più accessibili.

Il processo di equalizzazione di diverse forme di test, tuttavia, deve rispondere a rigorosi requisiti [3, 4, 5]:

- **Costrutto misurato**: Le diverse forme del test devono avere l'obiettivo di misurare lo stesso costrutto psicologico o abilità.
- **Affidabilità**: Ogni forma del test deve essere affidabile, nel senso che i punteggi ottenuti dai partecipanti dovrebbero essere consistenti e precisi. L'affidabilità è una misura della costanza dei risultati del test e della loro capacità di essere riproducibili.
- **Trasformazione di equating**: La trasformazione utilizzata per convertire i punteggi da una forma del test all'altra deve essere una funzione matematica precisa e invertibile, in modo che i punteggi possano essere convertiti senza perdita di informazioni.
- **Indipendenza dalla forma del test**: I punteggi dei candidati dovrebbero essere indipendenti dalla particolare forma del test che hanno completato. Ciò significa che non dovrebbero esserci vantaggi o svantaggi nell'affrontare una forma del test rispetto a un'altra.
- **Consistenza della funzione di equating**: La funzione di equating deve essere robusta e consistente, cioè deve fornire risultati equivalenti indipendentemente dalla scelta della popolazione da cui è derivata.

Esistono diverse funzioni di equalizzazione [6, 7, 8], che aumentano in complessità e precisione:

- **Identity equating (Equating di identità):**
  Non vi è alcuna equalizzazione, si assume che le distribuzioni dei punteggi differiscano solo a causa di rumore che non possiamo stimare. Questa è un'ipotesi molto forte, con un elevato potenziale di bias. Tuttavia, talvolta non si può applicare nessuna funzione di equating a causa di un campione troppo piccolo, quindi l'identità diventa il valore predefinito con dimensioni del campione insufficienti (ad esempio, inferiori a 30).
- **Mean equating**:
  Applica un aggiustamento costante a tutti i punteggi basato sulla differenza media tra le distribuzioni dei punteggi. Si stimano solo le medie, quindi i requisiti di dimensione del campione sono minimizzati (ad esempio, 30 o più), ma la possibilità di distorsioni nel punteggio è elevato, poiché l'aggiustamento medio potrebbe essere inappropriato per chi si colloca molto in basso o molto in alto nei punteggi.

- **Equipercentile equating (Equating equipercentile):**
  I passaggi fondamentali includono la stima delle funzioni di distribuzione cumulativa (CDF) per entrambe le forme, l'identificazione dei percentili corrispondenti, l'aggiustamento dei punteggi della forma ritenuta più difficile e l'uso di interpolazione e smussatura per ottenere una funzione di equipercentile continua. Questo approccio mira a minimizzare il bias, assicurando che i punteggi corrispondenti abbiano significato simile in termini di difficoltà a livello di percentili. L'equipercentile equating richiede dimensioni del campione più ampie, tipicamente nell'ordine di 200 o più, per ottenere stime accurate delle distribuzioni cumulative. In sintesi, l'equipercentile equating garantisce coerenza nelle interpretazioni dei punteggi su diverse forme di test.

- **Linear equating:**
  L'equating lineare utilizza un modello lineare con intercetta e pendenza per regolare i punteggi tra diverse forme di test. A differenza di approcci come l'equipercentile equating, che si focalizzano sulla difficoltà relativa, l'equating lineare consente l'aggiustamento dei punteggi lungo l'intera scala. Questo metodo può determinare incrementi o decrementi differenziali dei punteggi, consentendo una regolazione più flessibile rispetto ad altre tecniche. La stima della deviazione standard è coinvolta, riducendo il potenziale di bias, ma richiede dimensioni del campione più ampie, tipicamente oltre 100, per risultati affidabili. In sintesi, l'equating lineare offre una flessibilità maggiore nella regolazione dei punteggi su diverse forme di test.

- **Circle-arc equating**:
  La circle-arc equating è una funzione di equalizzazione non lineare che applica l'identify equating nelle code della distribuzione (per i punteggi troppo bassi o troppo alti) e il mean equating in corrispondeza dei valori medi. Questo permette di mantenere l'equivalenza nelle code, riducendo il bias potenziale, mentre si effettua un aggiustamento basato sulla media nelle regioni centrali della scala. L'approccio fornisce una soluzione flessibile e bilanciata per l'adattamento dei punteggi su diverse forme di test.

La nota 5 a pagina 12 della relazione del CISIA riporta un articolo mostra come l'approccio *circle-arc equating* sia più affidabile del *mean equating*, ma nonostante ciò, il CISIA ha adottato una rivisitazione del metodo equating, a livello di singola domanda, che presenta diverse criticità di seguito descritte.

2.1 Criticità del meccanismo di equalizzazione applicato dal CISIA

Di seguito viene decritto l'approccio di equalizzazione adottato dal CISIA (pagina 7 della relazione del 30/11/2023):

"In mean equating, Form X is considered to differ in difficulty from Form Y by a constant amount along the score scale. For example, under mean equating, if Form X is 2 points easier than Form Y for high-scoring examinees, it is also 2 points easier than Form Y for low-scoring examinees. [...] Thus, mean equating involves the addition of a constant (which might be negative) to all raw scores on Form X to find equated scores on Form Y."

Più precisamente, si definiscano due prove A e B del test e siano A e B, rispettivamente, le variabili aleatorie del punteggio non equalizzato delle prove A e B. Siano a *una realizzazione della variabile aleatoria A e* b *una realizzazione della variabile aleatoria B. Infine, siano* $\mu(A)$ *e* $\mu(B)$ *le medie, rispettivamente, delle variabili aleatorie A e B. In questo metodo vengono considerati uguali punteggi non equalizzati che differiscono ugualmente dalle relative medie, cioè* a *e* b *vengono considerati uguali se vale la seguente uguaglianza:*

$$a - \mu(A) = b - \mu(b) \tag{1}$$

Il CISIA stesso dichiara inoltre (sempre a pagina 7) che:

L'equalizzazione prevista per il TOLC-MED e il TOLC-VET implementa il metodo mean equating non a livello di prova, ma già a livello di singolo quesito, come si legge nell'allegato 2 del decreto ministeriale 1107 del 24 settembre 2022.

Nello specifico il meccanismo di equalizzazione prevede il calcolo del punteggio equalizzato come segue:

$$P_{equalizzato} = P_{non\_equalizzato} + C_{eq} \tag{2}$$

dove

$$C_{eq} = 50 - CdF \tag{3}$$

Il coefficiente di facilità (CdF) definito nell'allegato 2 al DM 1107 del 30 settembre 2022 è il seguente:

$$CdF = \sum_{i}^{50} \mu_i \tag{4}$$

dove μ_i rappresenta il punteggio medio ottenuto dai partecipanti per la domanda i-esima. ¡¡

Tuttavia, la letteratura in psicometria specifica nel settore dell'equalizzazione [9, 10]) è concorde nello stabilire che il metodo mean equating è applicato a livello di test, e non a livello di singola domanda, per le seguenti motivazioni:

1. Equalizzare a livello di singola domanda non tiene conto del fatto che la probabilità di una risposta corretta ad una domanda è una funzione sia del tratto sottostante misurato (ad esempio, abilità) e delle caratteristiche della domanda stessa (ad esempio, difficoltà) [9]. Quindi, a parità di preparazione, il punteggio di uno studente per un test dipende dalla difficoltà della domanda oltre che dalla sequenza di domande eseguite e dalla loro difficoltà.

2. La media dei punteggi delle domande dipende fortemente dal campione specifico di coloro che svolgono tali domande. Applicare il mean equating a livello di domanda può risultare in un'equalizzazione non generalizzabile oltre il gruppo specifico di coloro che hanno svolto il test per il processo di equalizzazione. Quindi, in assenza di un meccanismo di verificare dell'omogeneità dei campioni ai quali le domande sono somministrate, l'uso dei CdF calcolati nella sessione in aprile, ha introdotto un bias nelle valutazioni della sessione di luglio.

Si consideri un semplice esempio per comprendere il motivo per cui il mean equating si applica a livello di test e non di singola domanda. Si supponga di avere due test, Test A e Test B, ognuno contenente due domande. Questi test vengono somministrati allo stesso gruppo di studenti, e si vuole eguagliare i punteggi in modo che siano confrontabili. Di seguito è riportato il caso di mean equating a livello di domanda che a livello di test.

Test A:

- Domanda 1: $CdF = 0.8$ (relativamente facile)
- Domanda 2: $CdF = 0.2$ (relativamente difficile)
- **Media a livello di test** = (0.8 + 0.2) / 2 = 0.5

Test B:

- Domanda 1: $CdF = 0.6$ (moderatamente facile)
- Domanda 2: $CdF = 0.4$ (moderatamente difficile)
- **Media a livello di test** = (0.6 + 0.4) / 2 = 0.5

A livello di test, il mean equating suggerirebbe che il Test A e il Test B sono equivalenti in difficoltà perché i punteggi medi sono entrambi 0.5. Nessun aggiustamento sarebbe fatto perché le prestazioni medie sono uguali su entrambi i moduli.

Ora, si analizzi il caso del mean equating a livello di domanda, come proposto dal CISIA:

Per il Test A:

- Si aggiustano i punteggi della Domanda 1 e della Domanda 2 in modo che entrambe abbiano un punteggio medio di 0.5 (la media del test).

Per il Test B:

- Si aggiustano i punteggi della Domanda 1 e della Domanda 2 in modo che entrambe abbiano un punteggio medio di 0.5 (la media del test).

Il mean equating a livello di domanda suggerisce che entrambe le domande su entrambi i moduli sono di uguale difficoltà, il che non riflette accuratamente i dati originali. Il Test A originale aveva una domanda facile e una difficile, mentre il Modulo B aveva due domande di difficoltà moderata.
In questo modo, il mean equating a livello di singola domanda può introdurre un bias, poiché non tiene conto delle differenze specifiche tra le domande e assume che la difficoltà media delle domande sia rappresentativa delle differenze individuali tra le forme del test, specialmente nel caso in cui la difficoltà delle domande non è controllata. Per queste ragioni, gli studiosi in psicometria di solito impiegano metodi di eguagliamento più sofisticati, come l'Item Response Theory [9], progettato per tener conto delle complessità delle distribuzioni dei punteggi dei test e delle caratteristiche delle singole domande.

Riassumendo, il CISIA ha adottato il metodo di mean equating a livello di singola domanda, scelta non supportata dalla letteratura nel settore, e non ha implementato alcun meccanismo per garantire:

- **che le domande siano equilibrate in termini di difficoltà nelle diverse forme del test**;
- **uniformità a livello di campione per ogni domanda tra la sessione di aprile e quella di luglio**.

2.2 Criticità sul numero di test per quesito

La formulazione del meccanismo implementato dal CISIA presenta un'ulteriore criticità: il numero di studenti utilizzato per il calcolo del coefficiente di facilità (CdF) di ciascuna domanda deve essere uniforme, altrimenti si possono verificare distorsioni rilevanti, soprattutto in contesti con un elevato numero di partecipanti, come nel caso del test TOLC-MED.

Tuttavia, nonostante questa necessità di uniformità, i risultati presentati nelle tabelle 9, 12 e 13 della relazione del CISIA del 30/11/2023 (riportate in questo documento nelle Figure 2, 3, 4) dimostrano che tale condizione non è stata rispettata.

Per illustrare questo punto, prendiamo in considerazione il quesito MR2, come riportato nella Tabella 9 della relazione del CISIA. Tale quesito è stato somministrato a 3453 partecipanti, ottenendo un coefficiente di facilità pari a 0.14. Al contrario, il quesito T3 nella Tabella 12, somministrato a 3461 partecipanti, ha ottenuto un coefficiente di facilità pari a 0.72. Ciò implica uno scarto di 8 partecipanti in più per il quesito T3.

Supponiamo ora che il quesito MR2 fosse stato somministrato a ulteriori 8 partecipanti (per ottenere un equilibrio con il quesito T3), e si ipotizzino i due seguenti scenari:

- Gli 8 partecipanti in più rispondono correttamente alla domanda MR2, facendo variare il CdF da 0.142 a 0.144. Nonostante la variazione appa-

rentemente minima, su 50 quesiti questa differenza può tradursi in una variazione complessiva di 0.1 punti.

- Gli 8 partecipanti in più rispondono erroneamente alla domanda MR2, facendo variare il CdF da 0.142 a 0.141. Anche in questo caso, la differenza apparentemente minima potrebbe portare a una variazione complessiva di 0.15 punti per l'intero test.

È importante notare che, sebbene 0.1 possa sembrare un valore modesto, un incremento/decremento di 0.1 nel punteggio totale potrebbe tradursi in un cambiamento significativo nelle posizioni (200 posizioni in più o in meno) della graduatoria di Medicina e Chirurgia per l'anno 2023.

Inoltre, gli esempi riportati nelle Figure 1, 2, 3, 4 mostrano scenari in cui il numero di partecipanti a cui è stato somministrato un determinato quesito è pressoché simile (rispettivamente, 3456, 3453, 3461 e 3456). Tuttavia, considerando che la banca dati contiene 1.700 quesiti e che sono stati eseguiti circa 80.000 test ad aprile, ogni quesito è stato mediamente somministrato a 2.352 studenti ((80.000 $\times$ 50 (domande per test))/1700), un numero inferiore di circa 1.000 unità rispetto agli esempi riportati nella relazione del CISIA.

2.3 Criticità nella modalità di calibrazione

La modalità di equalizzazione definita dal CISIA e descritta nell'all. 2 al D.M. 1107 del 24/09/2022 prevede la presenza di un periodo di "calibrazione", coincidente con la prima sessione di erogazione dei test ad aprile 2023. Tale periodo di calibrazione è stato impiegato per valutare e fissare i coefficienti di facilità (CdF) di ciascuna domanda, calcolati come la media dei punteggio ottenuti in quella domanda durante la sessione di calibrazione.

Oltre alle già discusse criticità associate alla specifica interpretazione del metodo di mean equating adottato dal CISIA, la scelta di effettuare una fase di calibrazione per fissare in via definitiva la stima delle medie per domanda rappresenta un'ulteriore fonte di problemi.

Nello specifico, questa scelta presuppone che le caratteristiche della popolazione di studenti e le condizioni di somministrazione rimangano stabili nel tempo (come discusso anche in sezione 2.1), il che è perlomeno discutibile. Variazioni nelle coorti di studenti (il cui livello di preparazione può essere influenzato dal tempo avuto a disposizione) o anche modifiche nella percezione e nell'approccio al test (che possono consistere, ad esempio, nel sapere di avere o meno ulteriori tentativi a disposizione prima dell'inizio dell'anno accademico) possono influenzare significativamente la difficoltà percepita delle domande e, di conseguenza, i punteggi medi.

Inoltre, fissare i coefficienti di facilità (CdF), in base ai risultati di una sola sessione di calibrazione, non tiene conto della possibilità di variazioni casuali o sistematiche nelle prestazioni degli studenti. Eventuali anomalie o tendenze non rappresentative nella coorte di studenti della sessione di calibrazione possono

portare a una stima distorta dei CdF, che a sua volta influenzerebbe l'equità dei test successivi.

Il meccanismo adottato dal CISIA è noto, nell'ambito della metrologia[1], come "calibrazione statica" che *si applica quando la dimensione temporale non e' rilevante nelle misure*, come descritto in tutti i corsi universitari nell'ambito delle misure. Per esempio, il corso di "Instrumentation, Measurements, and Statistics" della Penn University afferma chiaramente che: *"Static calibration is performed when time is not relevant in the measurement"*[2].

Per mitigare i rischi associati alla calibrazione statica, sarebbe stato opportuno adottare una strategia di calibrazione dinamica. Questa avrebbe previsto una revisione periodica dei CdF basata sui dati raccolti in più sessioni, adattando la difficoltà delle domande alle effettive prestazioni degli studenti nel tempo. Questo approccio avrebbe contribuito a mantenere l'equità del test, considerando in modo più accurato le dinamiche delle prestazioni degli studenti nel corso del tempo.

2.4 Criticità nel calcolo dei coefficienti di facilità della relazione del CISIA del 30/11/2023

Nella sezione "Attribuzione del coefficiente di facilità ai quesiti" (pagg. 13-31) della relazione del CISIA del 30/11/2023, sono presentati esempi relativi al calcolo dei coefficienti di facilità delle domande, apparentemente **in base ai dati effettivamente acquisiti dal CISIA durante il periodo di calibrazione**.

In particolare, le Tabelle 8, 9, 12 e 13 riportano le prestazioni (in termini di risposte corrette, errate e omesse) di diversi cluster della popolazione usata per la calibrazione di quattro specifiche domande di test (indicate, rispettivamente, con B1, MR2, R3 e CF2). I valori relativi al numero di risposte corrette, errate e omesse sono riportati nel dettaglio per il campione afferente al Liceo Scientifico e in forma riassuntiva per il campione restante. Alla fine di ciascuna tabella, sono riportati i totali di risposte corrette, errate e omesse per il campione afferente al Liceo Scientifico, per il campione restante e per l'intera popolazione (costituita dall'unione del campione afferente al Liceo Scientifico e del campione restante). Naturalmente, ci si aspetta che il numero di risposte corrette per l'intera popolazione corrisponda alla somma del numero di risposte corrette per il campione afferente al Liceo Scientifico e del numero di risposte corrette per il campione restante (analogamente, ovviamente, per i numeri di risposte errate e omesse). In effetti, tale condizione è verificata per la Tabella 8; tuttavia, non è verificata per le Tabelle 9, 12 e 13.

Le Figure 1, 2, 3, 4 riportano gli estratti rilevanti delle tabelle in questione.

[1] La metrologia è la scienza e l'insieme di tecniche dedicate alla misurazione, e comprende lo sviluppo e l'utilizzo di strumenti di misura, nonché la definizione e l'interpretazione di standard di misura. L'obiettivo della metrologia è assicurare che le misure siano accurate, riproducibili e comparabili, indipendentemente dal luogo o dal tempo in cui vengono eseguite.

[2] https://www.me.psu.edu/cimbala/me345web_Fall_2014/Lectures/Errors_and_Calibration.pdf

1	2	3	4	5	6 - C	7 - C	6	7	8	9	10	11
prog	genere	tipo scuola	macroreg	livello istruzione	campione	% campione	iscritti totali	% iscritti totali	quesito	count_esatte	count_omesse	count_errate
CAMPIONE LICEO SCIENTIFICO					1936	56,02	38642	55,75		1013	569	354
RESTANTE CAMPIONE - ALTRE 240 RIGHE					1520	43,98	30675	44,25		741	408	371
CONTEGGIO ESATTE, NON DATE ED ERRATE TOTALE CAMPIONE - 3456 PARTECIPANTI										**1754**	**977**	**725**
CALCOLO COEFFICIENTE DI FACILITA' QUESITO B1 = (1754*1-0,25*725+977*0) /(1754+977+725)										**0,46**		

Figura 1: Estratto della Tabella 8 della relazione del CISIA del 30/11/2023.

1	2	3	4	5	6 - C	7 - C	6	7	8	9	10	11
prog	genere	tipo scuola	macroreg	livello istruzione	campione	% campione	iscritti totali	% iscritti totali	quesito	count_esatte	count_omesse	count_errate
CAMPIONE LICEO SCIENTIFICO					1930	55,89	38642	55,75		577	882	471
RESTANTE CAMPIONE - ALTRE 240 RIGHE					1523	44,11	30675	44,25		1177	95	254
CONTEGGIO ESATTE, NON DATE ED ERRATE TOTALE CAMPIONE - 3453 PARTECIPANTI										**712**	**1858**	**883**
CALCOLO COEFFICIENTE DI FACILITA' QUESITO MR2 = (712*1-0,25*883+1858*0) /(712+1858+883)										**0,14**		

Figura 2: Estratto della Tabella 9 della relazione del CISIA del 30/11/2023.

1	2	3	4	5	6 - C	7 - C	6	7	8	9	10	11
prog	genere	tipo scuola	macroreg	livello istruzione	campione	% campione	iscritti totali	% iscritti totali	quesito	count_esatte	count_omesse	count_errate
CAMPIONE LICEO SCIENTIFICO					1930	55,76	38642	55,75		1453	151	326
RESTANTE CAMPIONE - ALTRE 240 RIGHE					1531	44,24	30675	44,25		301	826	399
CONTEGGIO ESATTE, NON DATE ED ERRATE TOTALE CAMPIONE - 3461 PARTECIPANTI										**2622**	**252**	**587**
CALCOLO COEFFICIENTE DI FACILITA' QUESITO T3 = (2622*1-0,25*587+252*0) /(2622+587+252)										**0,72**		

Figura 3: Estratto della Tabella 12 della relazione del CISIA del 30/11/2023.

1	2	3	4	5	6 - C	7 - C	6	7	8	9	10	11
prog	genere	tipo scuola	macroreg	livello istruzione	campione	% campione	iscritti totali	% iscritti totali	quesito	count_esatte	count_omesse	count_errate
CAMPIONE LICEO SCIENTIFICO					1935	55,99	38642	55,75		996	599	340
RESTANTE CAMPIONE - ALTRE 240 RIGHE					1521	44,01	30675	44,25		758	378	385
CONTEGGIO ESATTE, NON DATE ED ERRATE TOTALE CAMPIONE - 3456 PARTECIPANTI										**1630**	**1138**	**688**
CALCOLO COEFFICIENTE DI FACILITA' QUESITO CF2 = (1630*1-0,25*688+1138*0) /(1630+1138+688)										**0,42**		

Figura 4: Estratto della Tabella 13 della relazione del CISIA del 30/11/2023.

È possibile notare come, nella Tabella 8 della relazione del CISIA, la somma dei conteggi delle risposte esatte per il campione Liceo Scientifico (1013) e il restante campione (741) è uguale al conteggio di risposte esatte in totale (1754 = 1013 + 741). Lo stesso vale per i conteggi delle risposte errate e omesse.

Tuttavia, analizzando la Tabella 9 della relazione del CISIA, la somma dei conteggi delle risposte esatte per il campione Liceo Scientifico (577) e il restante campione (1177) **non** è uguale al conteggio di risposte esatte in totale ($712 \neq 577 + 1177$). Lo stesso problema si verifica per i conteggi delle risposte errate e omesse, sia per la Tabella 9 che per le Tabelle 12 e 13 della relazione.

Desta preoccupazione il fatto che la non corrispondenza delle somme non sembri un errore di battitura, in quanto la discussione a contorno delle Tabelle prosegue utilizzando i valori riportati nelle tabelle. Non è chiaro, pertanto, se sono sbagliati i valori riportati per il campione afferente al Liceo Scientifico e/o per il campione restante, o se è sbagliato il valore totale per l'intera popolazione.

In quest'ultimo caso, ciò risulterebbe in una stima errata dei coefficienti

di facilità delle domande in esame. Ad esempio, analizzando la Tabella 9 della relazione, il coefficiente di facilità risultante considerando i valori riportati per il campione afferente al Liceo Scientifico e per il campione restante sarebbe uguale a:

$$CdF = \frac{(1 \cdot N_c) + (-0,25 \cdot N_e)}{N} = \frac{(1 \cdot 1754) + (-0,25 \cdot 725)}{3456} = 0,46 \quad (5)$$

Il valore risultante (0,46) è sensibilmente diverso da quello riportato nella relazione (0,14); se tale errore di calcolo si è effettivamente verificato, il procedimento di equalizzazione risulta gravemente invalidato.

Qualora, nella migliore delle ipotesi, i valori totali per l'intera popolazione riportati nelle Tabelle 9, 12 e e 13 della relazione fossero corretti, e di conseguenza i coefficienti di facilità risultassero validi, resterebbe da spiegare un'altra anomalia: il fatto che la somma dei valori parziali (campione Liceo Scientifico + campione restante) nelle Tabelle 9, 12 e 13 risulta sempre uguale ai valori per l'intera popolazione nella Tabella 8, sia per le risposte corrette, che omesse, che errate.

3 Numero dei quesiti componenti la banca dati

Il CISIA, a pagina 10 della relazione del 30/11/2023, dichiara che: "*Il numero di quesiti componenti la banca dati unitaria del TOLC-MED e TOLC-VET, nel primo anno di avvio 2023, è 1700. Questo numero nasce dalla necessità di soddisfare due esigenze da contemperare:*

- *costruire un numero sufficiente di prove diverse;*
- *ottenere misure statisticamente affidabili della difficoltà dei quesiti.*

È chiaro, infatti, che:

- *con pochi quesiti non è possibile produrre molte prove abbastanza diverse;*
- *con troppi quesiti non è possibile somministrare ogni singolo quesito un numero di volte sufficiente a garantire una misura affidabile della sua difficoltà.*

Il numero di quesiti componenti la banca dati unitaria del TOLC-MED e TOLC-VET, nel primo anno di avvio 2023, è 1700. Questo numero nasce dalla necessità di soddisfare due esigenze da contemperare: Ad esempio, una banca dati di 100 quesiti consente di produrre soltanto 2 prove senza sovrapposizioni e più in generale poche prove con pochi quesiti in comune. I partecipanti al test si trovano quindi nella situazione di rispondere troppo spesso agli stessi quesiti. È perciò evidente che tale banca dati non è sufficientemente numerosa. Al contrario, se anche si considerano 100 mila partecipanti, ciascuno dei quali affronta

50 quesiti, ottenendo quindi complessivamente 5 milioni di visualizzazioni di un quesito, una banca dati composta da 50 mila quesiti risulta troppo numerosa: mediamente un quesito viene visto 100 volte con il risultato che il suo coefficiente di facilità è troppo dipendente dalla risposta di un singolo candidato. In altre parole, in questo caso il suo coefficiente di facilità è una misura troppo approssimata e troppo sensibile al campione di partecipanti a cui il quesito è stato somministrato.

Malgrado tale considerazione possa essere corretta in linea di principio, è cruciale sottolineare che la scelta del numero di domande da utilizzare **non può e non deve essere effettuata basandosi esclusivamente su considerazioni qualitative**, come quelle presentate dal CISIA. In effetti, in statistica, esiste un metodo per stimare quantitativamente la dimensione di un campione (vale a dire, il numero di domande nella banca dati del CISIA in questo contesto) ed è universalmente noto come **sample size estimation**. Questo concetto rappresenta un concetto fondamentale ed elementare in statistica, il quale si basa sui seguenti fattori [11]:

- **Dimensione della popolazione**: numero totale delle persone appartenenti al gruppo che si sta analizzando. Nel caso del TOLC MED, il CISIA dichiara che sono stati somministrati 160.000 nelle due sessioni di aprile e luglio. Considerando che il calcolo dei coefficienti è fatto considerando solo la sessione di aprile, si può assumere che la popolazione è di 80.000 studenti.
- **Margine di errore**: percentuale che indica con quanta probabilità i coefficienti di difficoltà calcolati riflettano la difficoltà effettiva dei quesiti. Minore è il margine di errore, maggiore sarà la probabilità di ottenere la difficoltà esatta con un determinato livello di confidenza.
- **Livello di confidenza del campione**: percentuale che rivela quanto si possa essere sicuri che effettivamente l'errore sia nel margine fissato.

Fatte queste premesse, la stima della dimensione del campione si calcola tramite la seguente equazione:

$$dimensione\_campione = \frac{\frac{z^2 \cdot p \cdot (1-p)}{e^2}}{1 + (\frac{z^2 \cdot p \cdot (1-p)}{N \cdot e^2})} \tag{6}$$

dove:

- N = dimensione della popolazione
- e = margine di errore (in percentuale)
- z = punteggio z (Z-score)

Livello di confidenza	Punteggio z
80%	1.28
90%	1.65
95%	1.96
99%	2.58

Tabella 1: Punteggi Z

Il punteggio z indica di quante deviazioni standard una determinata proporzione dista dalla media. Per trovare il punteggio z corretto da utilizzare, si può fare riferimento alla Tabella 1.

Nel contesto del calcolo dei coefficienti di facilità del TOLC MED, prendendo in considerazione una popolazione di 80.000 studenti e riconoscendo l'importanza di tali punteggi per la graduatoria, con un livello di confidenza del 99% e un margine di errore dell'1%, l'applicazione della formula in Eq. 6 avrebbe richiesto che ogni singola domanda fosse somministrata a 13.776 studenti.

Tuttavia, un'analisi dettagliata, basata sulla banca dati dei quesiti del CISIA composta da 1.700 domande e sui circa 80.000 test eseguiti ad aprile, rivela che in media ogni quesito è stato somministrato a 2.352 studenti ((80.000 $\times$ 50 (domande per test))/1700). Questo numero è largamente insufficiente per ottenere un errore dell'1% con un livello di confidenza del 99% sull'accuratezza del calcolo.

Da notare inoltre che se si desidera un livello di confidenza più elevato (ad esempio, 99.9%) e un margine di errore dello 0.1%, il numero richiesto aumenta da 13.776 a 20.221. Questi dati evidenziano la discrepanza tra la dimensione del campione effettivamente utilizzata dal CISIA e quella necessaria per garantire una precisione statistica adeguata, per poter affermare che il CdF calcolato rappresenti la reale difficoltà dei quesiti.

Pertanto l'approccio utilizzato dal CISIA per stabilire il numero di quesiti da somministrare non è supportato da alcuna evidenza statistica.

4 Criteri e modalità di individuazione dei quesiti

Il CISIA dichiara, a pagina 11, quanto segue:

"In letteratura si rileva come nella maggior parte delle situazioni per l'individuazione dei quesiti da sottoporre ai candidati è sufficiente utilizzare l'equivalent group design, detto anche random group design: a ciascun partecipante al test viene somministrata casualmente una delle possibili diverse prove. È ben noto che uno svantaggio dell'equivalent group design riguarda la dimensione dei campioni, che devono essere sufficientemente numerosi."

Tuttavia, diversi studi evidenziano che il random group design può portare a risultati inattesi e distorti [12], soprattutto in assenza di un controllo accurato della difficoltà delle domande. Non solo il livello di complessità medio dei singoli test [13, 14], ma anche l'ordine della difficoltà delle domande nei test [15], sono parametri da considerare attentamente. La principale ragione per tenere in considerazione questi aspetti è l'impatto psicologico sugli studenti: affrontare una prova più difficile può generare sensazioni di scoraggiamento, ansia o pressione, influenzando negativamente le loro prestazioni reali. Gli studiosi Ayana et al. in [15] affermano: *"potential psychological effects stemming from the order of difficult questions are particularly important in tests in which participants cannot access later questions until having answered the previous ones. A common practice by many academics is to simply randomize the order of questions or order questions in the order in which particular topics were explained during a course. However, these practices do not necessarily take into account potential psychological effects, if they do not explicitly incorporate the potential concern in the test design"*. In altre parole, gli effetti psicologici dovuti alla mancanza di controllo della difficoltà dei test, specialmente in quei test in cui non è possibile accedere alle domande successive sin dall'inizio (come nel caso del TOLC-MED), portano a un sistema iniquo, poiché il confronto tra i risultati di studenti che affrontano test con livelli di difficoltà diversi è distorto e non rappresentativo delle loro reali abilità.

Anche in presenza di un meccanismo di equalizzazione, è difficile garantire un sistema completamente equo. Il processo di equalizzazione non compensa gli effetti psicologici causati da differenze nei livelli di difficoltà.

In conclusione, l'assenza di controllo della difficoltà delle domande nei vari test da parte del CISIA compromette la validità del metodo, poiché non è accuratamente progettato per misurare in modo efficace le abilità degli studenti. La coerenza nel livello di difficoltà dei test è un elemento essenziale per garantire l'obiettivo di valutare le abilità degli studenti con precisione, senza essere influenzati da fattori esterni come la difficoltà oggettiva di una specifica prova.

5 Conclusioni

In conclusione, l'analisi critica del sistema TOLC-MED evidenzia molteplici lacune che compromettono la sua validità e equità. Le principali criticità riguardano il meccanismo di equalizzazione adottato dal CISIA, il numero insufficiente di test per ciascuna domanda, la modalità di calibrazione statica e la mancanza di criteri adeguati per l'individuazione e il controllo della difficoltà delle domande.

- In primo luogo, l'approccio di equalizzazione a livello di singola domanda, basato sul metodo mean equating, è in contrasto con i principi fondamentali della psicometria, che suggeriscono l'applicazione del mean equating a livello di test. Questa scelta introduce distorsioni nei punteggi degli stu-

denti, ignorando le complessità legate alla variazione della difficoltà delle domande e alla diversità dei campioni di studenti.

- Gli esempi di calcolo del coefficiente di facilità per alcuni quesiti presentati nella relazione (Tabelle 8, 9 e 12) sono errati e portano a valori differenti dei vari coefficienti.

- La modalità di calibrazione statica adottata (usando solo i coefficienti di facilità calcolati nella sessione di aprile), senza una revisione periodica basata su dati raccolti in diverse sessioni, introduce il rischio di distorsioni nella stima dei coefficienti di facilità a causa di variazioni casuali o sistematiche nelle prestazioni degli studenti.

- Infine, la mancanza di criteri e controllo sulla difficoltà delle domande nei test compromette l'equità del sistema, poiché gli effetti psicologici derivanti da differenze nei livelli di difficoltà non vengono adeguatamente considerati.

Tuttavia, la criticità maggiore che, a parere dello scrivente, invalida l'intero test è quella relativa al numero di studenti usato per calcolare il coefficiente di facilità di ogni domanda: le domande sono state somministrate ad un numero di studenti notevolmente più basso (in media 2.352 studenti) rispetto a quanto richiesto (almeno 13.776 studenti) per poter affermare (con una probabilità del 99%) che la stima del coefficiente di facilità fosse affidabile (entro un margine di errore dell'1%).

Questo mette in luce un'applicazione inadeguata del concetto di sample size estimation, con conseguente compromissione della precisione statistica e dell'affidabilità dei coefficienti di facilità calcolati.

In sintesi, il sistema TOLC-MED si presenta come inadeguato, inaccurato, iniquo e progettato superficialmente, trascurando i principi fondamentali di statistica, metrologia e psicometria. Sono necessarie revisioni significative per garantire un processo di valutazione equo, accurato e affidabile per gli studenti partecipanti.

6 Breve descrizione delle attività dell'Ing. Carmelo Pino

Carmelo Pino è membro del team Artificial Intelligence for Modeling and Predictive Reliability di STMicroelectronics (ADG - R&D Power and Discretes) in qualita' di Advanced Research Senior Engineer. Ha lavorato come assistente di ricerca presso l'Università di Catania, dove ha completato il suo dottorato di ricerca. Dal 2014 ad oggi è membro del Laboratorio di Pattern Recognition e Computer Vision dell'Università di Catania e dal 2020 al 2022 è stato assistente di ricerca presso l'INAF (Istituto Nazionale di Astrofisica), Catania, Italia. I suoi interessi di ricerca includono le aree dell'intelligenza artificiale applicata e l'analisi dei dati applicati al medical data processing, temporal series analysis. In tali ambiti ha pubblicato oltre 60 articoli scientifici in riviste e conferenze di alto prestigio, e ha ricevuto oltre 500 citazioni. Il suo lavoro "Semantic Segmentation of radio-astronomical images" ha ricevuto il premio come miglior articolo all'International Workshop on Artificial Intelligence and Pattern Recognition, 2021.

Riferimenti bibliografici

[1] William Stout. Psychometrics: From practice to theory and back: 15 years of nonparametric multidimensional irt, dif/test equity, and skills diagnostic assessment. *Psychometrika*, 67:485–518, 2002.

[2] Joseph Ryan and Frank Brockmann. A practitioner's introduction to equating with primers on classical test theory and item response theory. *Council of Chief State School Officers*, 2009.

[3] Neil J Dorans, Tim P Moses, and Daniel R Eignor. Principles and practices of test score equating. *ETS Research Report Series*, 2010(2):i–41, 2010.

[4] Anthony D Albano and Marie Wiberg. Linking with external covariates: Examining accuracy by anchor type, test length, ability difference, and sample size. *Applied psychological measurement*, 43(8):597–610, 2019.

[5] Deborah J Harris and Jill D Crouse. A study of criteria used in equating. *Applied Measurement in Education*, 6(3):195–240, 1993.

[6] Anthony D Albano. A general linear method for equating with small samples. *Journal of Educational Measurement*, 52(1):55–69, 2015.

[7] Anthony D Albano, Theodore J Christ, and Liuhan Cai. Evaluating equating in progress monitoring measures using multilevel modeling. *Measurement: Interdisciplinary Research and Perspectives*, 16(3):168–180, 2018.

[8] Jorge González and Marie Wiberg. Applying test equating methods. *Cham: Springer International Publishing*, 2017.

[9] Rafael Jaime De Ayala. *The theory and practice of item response theory.* Guilford Publications, 2013.

[10] Sevilay Kilmen and Nukhet Demirtasli. Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia - Social and Behavioral Sciences*, 46:130–134, 2012. 4th WORLD CONFERENCE ON EDUCATIONAL SCIENCES (WCES-2012) 02-05 February 2012 Barcelona, Spain.

[11] Bernard Rosner. *Fundamentals of biostatistics /.* 8th edition. edition.

[12] Per-Erik Lyrén and Ronald K Hambleton. Systematic equating error with the randomly-equivalent groups design.

[13] M. Spencer, A. F. Gilmour, A. C. Miller, A. M. Emerson, N. M. Saha, and L. E. Cutting. Understanding the Influence of Text Complexity and Question Type on Reading Outcomes. *Read Writ*, 32(3):603–637, Mar 2019.

[14] Asghar Salimi, Soghra Dadaspour, and Hassan Asadollahfam. The effect of task complexity on efl learners' written performance. *Procedia - Social and Behavioral Sciences*, 29:1390–1399, 2011.

[15] Lina Anaya, Nagore Iriberri, Pedro Rey-Biel, and Gema Zamarro. Understanding performance in test taking: The role of question difficulty order. *Economics of Education Review*, 90:102293, 2022.