ChatGPT

Predizione dell'Umore: Conformal Prediction, FDA in RKHS e Qualità del Survey

1. Conformal Prediction per la Predizione dell'Umore da Questionari

Le tecniche di *conformal prediction* (CP) stanno trovando applicazione nella **predizione di stati d'umore e di salute mentale** per fornire stime incertezza affidabili. In queste applicazioni di *machine learning* in ambito psicologico, CP viene usato tipicamente in modalità *batch/offline* con approccio **inductive (split) conformal**: il modello viene addestrato su un training set e poi calibrato su un calibration set separato, garantendo copertura predittiva valida sotto l'assunzione di exchangeability (i.i.d.) [1]. Ad esempio, Li e Zhou (2024) hanno proposto *Conformal Depression Prediction (CDP)* per quantificare l'incertezza nella stima della gravità depressiva da video facciali [2] [1]. In questo studio (dominio: *affective computing*), un regressor pre-addestrato fornisce un punteggio (es. punteggio BDI-II, 0–63) e CP viene utilizzato per costruire **intervalli di previsione** attorno a tale punteggio, garantendo copertura marginale al livello desiderato (es. 90%) [3] [4]. Gli autori adottano un **inductive conformal predictor**: utilizzano un sottoinsieme di dati come set di calibrazione per calcolare i *conformal scores* e costruire intervalli [1]. Hanno inoltre introdotto un'estensione (*CDP-ACC*) per approssimare la *conditional coverage*, restringendo la larghezza degli intervalli per input con incertezza minore [5] [6].

Metriche di valutazione: Nei compiti di regressione sull'umore, le prestazioni dei conformal predictors vengono misurate tramite la *Prediction Interval Coverage Probability* (PICP) – percentuale di casi in cui il vero valore cade nell'intervallo predetto – e la *Mean Prediction Interval Width* (MPIW) – ampiezza media degli intervalli [7]. Li e Zhou ad esempio riportano che con miscoverage 0.1 (confidence 90%) i loro intervalli CP ottengono copertura effettiva ~90% e larghezze inferiori rispetto al conformal standard grazie all'algoritmo adattivo proposto [7] [8]. Nel caso di **classificazione dell'umore** (es. riconoscimento di emozioni), CP fornisce *prediction sets* (insiemi di etichette possibili) invece di un singolo output, garantendo che la *label* reale sia inclusa con probabilità $\geq$ $1-\alpha$. Roohi et al. (2024) hanno integrato CP in un modello NLP per riconoscere emozioni in conversazione, migliorando la *calibrazione* delle probabilità e riducendo bias di classe [9]. In particolare, evidenziano che i modelli di riconoscimento emotivo tendono a sovrapredire l'emozione *"neutral"*; l'uso di CP produce insiemi di emozioni plausibili (es. {tristezza, neutral, gioia}) anziché una singola previsione, ottenendo copertura marginale controllata e previsioni più affidabili [10]. Propongono un variante denominata *class spectrum conformation* per garantire copertura più uniforme tra classi emotive, mitigando il bias intrinseco dei classificatori [10].

Tipologie di conformal predictor: Nella letteratura 2015–2025 sull'umore, la quasi totalità delle applicazioni adotta CP in forma *induttiva batch*. Il metodo *transductive* originale (ricalcolo per ogni test point) è concettualmente presente ma impraticabile su larga scala, mentre approcci *online* CP (aggiornamento continuo con stream di dati) non compaiono nei lavori esaminati, dato che gli studi sul campo tendono a lavorare su dataset raccolti interamente prima dell'analisi. In sintesi, i conformal predictors per questionari d'umore sfruttano una fase di calibrazione separata e forniscono **intervalli di confidenza o insiemi di etichette** per ogni previsione, con garanzie formali di accuratezza. Ciò consente, ad esempio, di predire un punteggio di depressione con il suo intervallo al 95%, oppure di

indicare un set di possibili stati d'animo correnti di un utente con livello di confidenza prefissato, migliorando la fiducia e l'interpretabilità nell'uso clinico [11] [9].

2. Functional Data Analysis in RKHS per Dati di Survey Psicologici

In ambito statistico, i **metodi di Functional Data Analysis (FDA)** permettono di trattare misurazioni ripetute o profili di risposta come *curve* o funzioni in uno spazio di Hilbert. Negli ultimi anni, sono stati sviluppati approcci per gestire **dati funzionali non tradizionali** – ad esempio dati ordinali Likert o binari da questionari psicologici – utilizzando spazi di Hilbert con *kernel* riproducibili (RKHS). Questi metodi mappano i dati discreti a funzioni latenti continue, su cui applicare tecniche di smoothing, PCA funzionale, regressione, ecc., mantenendo la natura ordinale o categorica dei dati originali.

Un lavoro di riferimento è quello di Dey et al. (2024), che ha introdotto una **FPCA (Functional Principal Component Analysis) per dati funzionali ordinali, binari e troncati** in modo unificato [12]. Questo studio, motivato da un dataset EMA (Ecological Momentary Assessment) del *Family Study of Mood Disorders* (497 partecipanti), considera ad esempio le valutazioni di *umore* 4 volte al giorno su scala Likert 1–7 (felice ↔ triste) trattandole come **curve ordinali nel tempo** [13]. Gli autori adottano un modello semiparametrico basato su *Gaussian copula* latente: si assume che dietro alle risposte ordinali ci sia un processo latente continuo, e si stimano covarianza e componenti principali di questo processo latente usando metodi in RKHS (Kendall's tau bridging) [12] [14]. In pratica, ciò consente di effettuare un'**analisi funzionale** sulle traiettorie temporali di mood (o altri item di survey: dolore sì/no, eventi, ecc.), combinando dati di natura differente (continui, binari, ordinali) in un unico framework [12] [14]. Il risultato è la possibilità di estrarre componenti principali e pattern di variazione dell'umore nel tempo, anche se le misure sono ordinali e sparse. Questo metodo è implementato in un software R fornito dagli autori [14]. Le analisi su dati reali hanno evidenziato, ad esempio, differenze significative nelle **traiettorie di umore giornaliere** tra pazienti con Disturbo Depressivo Maggiore e Bipolare [14], dimostrando l'utilità di FDA in contesto psicopatologico.

Altri approcci in RKHS per dati psicologici includono metodi per dati categoriali puramente qualitativi. Preda et al. (2021) hanno esteso l'analisi funzionale a **processi categorici a stati finiti** (sequenze di categorie nel tempo) introducendo la cosiddetta *categorical functional data analysis*. La loro metodologia, implementata nel package R `cfda` [15] [16], approssima le traiettorie categoriali mediante un insieme di funzioni base (es. basi B-spline o Fourier) e trova rappresentazioni ottimali nello spazio funzionale, analogamente a come l'analisi per corrispondenze multiple tratta dati categoriali statici [17]. Questo consente di ridurre la dimensionalità e visualizzare pattern anche su dati come *sequence di stati emotivi* (ad es., sequenza di emozioni sperimentate durante la giornata, ciascuna in {gioia, tristezza, rabbia,...}). Sebbene le scale Likert abbiano natura ordinale, tecniche simili possono essere applicate: si può trattare una risposta Likert come stato categoriale oppure come valore numerico associato a una funzione base. In quest'ottica, l'RKHS fornisce un formalismo in cui **incorporare la natura ordinale**: ad esempio, Habing (2017) ha definito *ordinal smoothing splines* con kernel appositi che rispettano l'ordinamento intrinseco delle categorie [18]. Questo permette di usare splines e metodi di smoothing sulle variabili ordinali (come item Likert) senza violare la loro scala (invarianti a trasformazioni monotone) [19] [20]. Tali approcci hanno il beneficio di non dover trattare forzatamente l'item Likert come numerico continuo (rischiando distorsioni) né come nominale puro (perdendo informazione sull'ordine).

R packages e strumenti: La comunità R offre diversi pacchetti per FDA e RKHS applicati a dati psicologici. Oltre al citato `cfda` [15], il task view CRAN su Functional Data Analysis elenca: `fda` (strumenti base FDA), `fda.usc` (analisi esplorative, classificazione e regressione con covariate funzionali, anche metodi di *depth* e outlier detection), e `refund` (regressione funzionale penalizzata,

PCA funzionale multivariata, incluse estensioni a *exponential family* per gestire esiti binari o count) [21] [22]. In particolare, i metodi di Goldsmith et al. (2015) e Wrobel et al. (2019) integrati in `refund` consentono **regressioni funzione-su-scalare e PCA** per dati non Gaussiani (es. esiti binari/ordinali) mediante link logit/probit e registrazione delle curve [23]. Inoltre, il pacchetto `conformalInference.fd` [24] offre procedure conformali per risposte funzionali, e il pacchetto `GPFDA` implementa processi gaussiani funzionali, utili per modellare curve di Likert come realizzazioni di GP. In generale, un workflow in R per *survey* psicologici longitudinali potrebbe includere: (a) rappresentare le risposte seriali degli item come classi funzionali (usando ad es. `funData`/`tf` per creare oggetti funzionali da dati discreti); (b) applicare smoothing rispettando la natura discreta (ad es. trasformata di probit per item ordinali [25] [26]); (c) eseguire PCA funzionale o clustering; (d) usare regressione in RKHS (come spline ordinali o kernel methods) per predire variabili di outcome. Ad esempio, Xie (2021) ha esplorato un *function-on-scalar regression* con link di copula gaussiana in RKHS, evidenziando la flessibilità di combinare FDA e kernel methods per dati psicometrici complessi [27].

In sintesi, l'uso di FDA in RKHS consente di **catturare l'andamento temporale e la struttura interna** dei dati di questionari (p.es. variazioni di umore diurno, profili di risposta su scale multi-item) trattandoli come funzioni, anche se i dati originali sono ordinali o categoriali. I pacchetti R come `fda.usc`, `refund` e `cfda` forniscono strumenti pronti per implementare tali analisi, mentre lavori recenti offrono codice open-source (ad es. *R-package* del metodo di Dey et al. 2024 [14]) per integrare dati Likert e binary in un'analisi funzionale unica.

3. Effetto di Risposte Casuali/Rumorose sulla Predizione dell'Umore

Le *risposte casuali o disattente* nei questionari – note anche come *careless responding* o *insufficient effort responding* – rappresentano un grave rischio per la qualità dei dati e la validità delle predizioni sull'umore. La letteratura evidenzia che persino una **piccola percentuale di risposte casuali (anche ~5–10%) può introdurre distorsioni sostanziali** nelle analisi [28] [29]. Ad esempio, simulazioni su dati di personalità mostrano che con appena ~10% di soggetti che rispondono in modo casuale, l'adattamento dei modelli fattoriali peggiora drasticamente e compaiono correlazioni spurie tra costrutti [28] [29]. In ambito predizione dell'umore, questo si traduce in modelli meno accurati e più "rumorosi": i pattern veri possono essere mascherati dal rumore, portando a errori di classificazione o a intervalli di predizione inutilmente larghi. Come sintetizzato da Curran et al., **dati spazzatura in ingresso producono conclusioni spazzatura in uscita** – anche pochi casi di scarsa qualità diminuiscono l'affidabilità dei risultati [30].

Diversi studi tra 2015 e 2025 hanno investigato l'impatto e soprattutto sviluppato **strategie di rilevamento e mitigazione** di queste risposte inattendibili:

- **Indicatori statistici di careless responding:** Si sono proposti numerosi indici per identificare ex-post i soggetti "sospetti". Ad esempio, **Intra-Individual Response Variability (IIRV)** o **Inter-Item Standard Deviation (ISD)** misura la variabilità delle risposte di un individuo: valori insolitamente bassi indicano che il rispondente ha dato quasi sempre lo stesso punteggio (p.es. ha messo "3" a tutte le domande), un pattern tipico di risposte poco attente [31]. Marjanovic et al. (2015) hanno mostrato che ISD discrimina bene tra risposte coscienziose vs random [31]. Altri indici comuni includono la lunghezza della stringa più lunga di risposte identiche (*longstring* index), la correlazione pari-dispari (coerenza tra item dispari e pari di una scala), e la distanza di Mahalanobis (per rilevare pattern di risposte multivariati anomali) [32] [33]. Il pacchetto R `careless` (Yentes & Wilhelm, 2018) implementa molte di queste metriche e facilita la scansione di dataset survey per *risposte insufficientemente impegnate* [32]. Ad esempio,

applicando tali indici ai dati di un questionario sull'umore, si potrebbero filtrare quei partecipanti che mostrano variabilità interna nulla o schemi incoerenti (es.: valutano di essere "molto felici" e "molto tristi" simultaneamente, il che potrebbe indicare scarsa attenzione) [34] [35].

- **Effetti sul modello predittivo:** Goldammer et al. (2020) hanno quantificato l'impatto delle risposte casuali sulle proprietà psicometriche e sulle predizioni. Hanno riscontrato che la presenza di *careless respondents* tende a **abbassare le consistenze interne** (Cronbach α) delle scale e a **biasare i coefficienti** in modelli di regressione [36]. In contesti di predizione dell'umore, ciò può significare pesi erratici attribuiti a certi item (per via di pattern di risposta anomali) o soglie di classificazione subottimali. La rimozione dei soggetti inattendibili porta in genere a un netto miglioramento: *fit* dei modelli più aderente, correlazioni tra variabili coerenti con le attese teoriche, maggiore potere predittivo [34] [35]. Un esempio concreto è fornito da Jaso et al. (2022) nello scenario EMA: analizzando ~18.000 auto-report emotivi in tempo reale, hanno individuato criteri semplici ma efficaci per flaggare un *assessment* come potenzialmente "careless" – **tempo di completamento eccessivamente rapido** ($\leq$1 secondo per item), **varianza delle risposte all'interno del questionario ~0** (es. dare quasi lo stesso voto a tutte le domande, SD $\leq$ 5 su una scala 0–100) o **uso ripetuto della stessa opzione** ($\geq$60% degli item con la medesima risposta) [37]. Eliminando le somministrazioni che soddisfacevano questi criteri, sono scomparse *correlazioni implausibili* come punteggi alti contemporaneamente su "rilassato" e "ansioso" (antitetici) [34]. Inoltre, identificando i *careless responses* si possono identificare anche i *careless responders* cronici (chi fallisce ripetutamente i criteri), candidati per essere esclusi interamente dalle analisi [38]. Questa strategia migliora l'accuratezza della predizione dell'umore, poiché il modello si basa solo su dati affidabili.

- **Metodi machine learning per detection:** Un filone recente utilizza algoritmi supervisati per riconoscere pattern di risposte random. Schroeders et al. (2022) hanno addestrato modelli di *gradient-boosted trees* su dataset etichettati (in cui alcuni partecipanti erano istruiti a rispondere casualmente) per distinguere questi ultimi dai rispondenti attenti. I risultati indicano un'**accuratezza elevata** nel classificare i careless responders, superiore ai singoli indici tradizionali presi isolatamente [39]. Questo approccio data-driven combina molteplici caratteristiche delle risposte (tempo, variabilità, coerenza su item correlati, ecc.) in modo non lineare, fornendo uno strumento potente per pulire il dataset prima di costruire il modello di predizione dell'umore. Tali tecniche sono complementari agli indici semplici: possono essere implementate in R (ad es. usando `xgboost` o `caret`) addestrando un classificatore sulle metriche di `careless` come input, se si dispone di un qualche gold standard o di iniezione simulata di risposte rumorose.

Mitigazione e best practice: La mitigazione principale è **escludere o correggere** le risposte identificativamente casuali. La review di Ward e Meade (2023) raccomanda di *screenare* i dati e riportare quanti casi sono stati rimossi per scarsa qualità [40]. In alcuni casi, se il tasso di careless responding è basso (es. <5%), la semplice rimozione dei soggetti incriminati è la soluzione più pulita. Se invece è più elevato, si possono considerare approcci robusti: p.es. utilizzare modelli a effetti misti che riducono l'influenza di outlier, oppure ponderare le osservazioni in base a un indice di qualità (down-weight di soggetti con indici di careless elevati). In studi longitudinali EMA, Jaso et al. propongono addirittura di implementare **monitoraggio in tempo reale**: con il package R `EMAeval` sviluppato dal loro team, i ricercatori possono impostare alert durante la raccolta dati (via smartphone) quando un partecipante risponde troppo velocemente o in modo piatto, permettendo di inviargli un messaggio di richiamo sul momento [41]. Questa interazione in tempo reale può prevenire dati inutilizzabili, migliorando la qualità prima ancora dell'analisi. In definitiva, la letteratura concorda sul fatto che la robustezza di qualunque modello predittivo dell'umore dipende fortemente dalla **qualità del dataset**: rilevare e mitigare le

risposte casuali è un passo cruciale, ottenuto combinando *design* accurato (vedi sezione 5) e *cleaning* data-driven post-raccolta [36] [34].

4. Questionari Psicometrici Validati per la Predizione dell'Umore

La ricerca sulla predizione dell'umore fa largo uso di **questionari standardizzati** e validati, che forniscono misure affidabili di stati affettivi, sintomi e tratti psicologici correlati all'umore. Questi strumenti, sviluppati in ambito psicometrico, garantiscono *validità di costrutto* e *affidabilità* elevate, fungendo sia da variabili indipendenti (feature) sia come variabili dipendenti (outcome) nei modelli predittivi. Ecco alcuni esempi chiave, con enfasi su struttura, tipo di variabili e proprietà psicometriche:

- **PANAS (Positive and Negative Affect Schedule):** È forse la scala di *affect* più utilizzata al mondo [42]. Strutturalmente consta di 20 aggettivi emozionali (es. *entusiasta, irritabile, ecc.*); il rispondente valuta in che misura li sta provando su una scala Likert a 5 punti (1 = per niente, 5 = estremamente) [43] [44]. Il PANAS produce due punteggi sommativi – **Affetto Positivo (PA)** e **Affetto Negativo (NA)** – considerati dimensioni indipendenti dell'esperienza affettiva [45]. Le variabili item sono ordinali (Likert), ma i punteggi finali (somma di 10 item ciascuno, range 10–50) vengono trattati come quantitativi a intervallo. **Validità e affidabilità:** Il PANAS originale (Watson et al., 1988) e successive traduzioni hanno mostrato eccellente consistenza interna, con **Cronbach α ~0.85–0.90 per PA e ~0.84–0.87 per NA** in campioni adulti. Studi recenti confermano la struttura a due fattori invariata in diverse popolazioni e contesti (generale, clinico, differenti culture) [46]. Ad esempio, Mor et al. (2020) hanno validato la versione online spagnola in un campione clinico (depressione, ansia): l'analisi fattoriale confermativa ha replicato i due fattori indipendenti, e la scala ha mostrato **ottima validità convergente** (correlando come atteso con misure di soddisfazione di vita e sintomi depressivi), **buona consistenza interna** ($\alpha \approx 0.9$) e **sensibilità al cambiamento** in interventi clinici [47] [46]. Nell'ambito predittivo, il PANAS viene usato sia come **variabile target** (punteggio PA o NA predetto da altre misure, p.es. dati da wearable [48]) sia come **fonte di feature** (gli item o i punteggi entrano in modelli che predicono esiti come rischio depressivo, benessere, ecc.). Essendo una misura di *stato emotivo corrente* (o di *tratto generale* a seconda delle istruzioni), il PANAS permette di standardizzare l'output di "umore" su una scala ben definita e con proprietà psicometriche note, facilitando il confronto tra studi.

- **BDI-II (Beck Depression Inventory-II):** È uno strumento classico per misurare la **sintomatologia depressiva** e viene spesso impiegato come outcome quantitativo nelle predizioni di stati depressivi. Struttura: 21 item, ciascuno composto da 4 affermazioni ordinate per gravità (punteggio 0–3); il rispondente sceglie quella che meglio descrive il proprio stato nell'ultima settimana. Gli item coprono sintomi cognitivi, affettivi e somatici della depressione (es. umore triste, senso di colpa, alterazioni del sonno) [49] [50]. La variabile totale (sommatoria 0–63) è trattata come quantitativa continua; spesso viene anche categorizzata in classi cliniche (0–13 assente/minima, 14–19 lieve, 20–28 moderata, $\geq$29 severa). **Validità psicometrica:** Il BDI-II è altamente affidabile e valido. In una recente valutazione su adolescenti (Lee et al., 2017), **Cronbach α = 0.89** per il punteggio totale [51], in linea con meta-analisi che riportano α intorno a 0.90–0.92 in adulti [52]. La validità convergente è eccellente: ad esempio, BDI-II e PHQ-9 (altro strumento depressivo) correlano ~0.75–0.80 [53]. Le analisi fattoriali tipicamente trovano due fattori (cognitivo-affettivo vs somatico) o tre fattori, ma ciò non inficia l'uso del punteggio totale che possiede **significato clinico** ampiamente verificato [54] [55]. In predizione dell'umore, il BDI può essere sia target (prevedere il punteggio di depressione di una persona in base ad altri dati, con approcci di regressione/interval prediction come in CDP [56] [6]) sia input (ad es. usare item o sottopunteggi BDI come predittori di drop-out o di outcome di trattamento). L'uso di BDI-II garantisce che il costrutto "depressione" sia misurato in modo standardizzato, aumentando la

validità esterna dei modelli: ad esempio, un modello che predice BDI-II≥30 (depressione grave) può essere utilmente confrontato con criteri clinici.

- **Altre scale di umore e affetto:** Numerosi altri questionari validati vengono usati a seconda del focus specifico. *Esempi:* **Profile of Mood States (POMS)** – 65 item Likert 0–4, fornisce 6 sub-scale (Tensione, Depressione, Rabbia, Vigore, Stanchezza, Confusione) e un indice globale di "disturbo dell'umore"; è impiegato specialmente in psicologia dello sport e medicina, con buone validità e α spesso >0.90 per il totale. **DASS-21** (Depression Anxiety Stress Scales) – 21 item Likert 0–3, tre scale da 7 item ciascuna; usato per monitorare stress e affetti negativi in popolazione generale e clinica (α≈0.88 per Depression). **PHQ-9 e GAD-7** – brevi scale diagnostiche per depressione e ansia (9 e 7 item rispettivamente, basate su criteri DSM); spesso utilizzate in studi su larga scala per la loro brevità, con sensibilità/specificità clinica stabilite e α ~0.8–0.9. **PANAS-X** – versione estesa del PANAS con 60 item coprendo affetti specifici (paura, ostilità, timidezza, ecc.), utile se si vogliono predire sfumature emotive particolari; validata da Watson & Clark (1994), mantiene struttura fattoriale gerarchica (fattore PA e NA sovra-ordinati). Tutti questi strumenti condividono l'uso di **item Likert o simili ordinali** aggregati in punteggi *quantitativi*, standardizzati tramite studi normativi e con robuste evidenze di validità.

In fase di progettazione di studi predittivi, scegliere questionari psicometrici validati assicura che le variabili d'interesse (umore, affetto, sintomi) siano *misurate con poco errore* e veramente rappresentative del costrutto teorico. Ad esempio, se l'obiettivo è predire il "miglioramento dell'umore", utilizzare la differenza di punteggio PANAS-PA pre/post intervento (che è affidabile e sensibile al cambiamento [47]) sarà preferibile rispetto a domande ad-hoc non validate. Inoltre, le **proprietà psicometriche** note (α, fattori) permettono di applicare tecniche appropriate: p.es., sapendo che PANAS produce due variabili correlate ma distinte, un modello potrebbe dover predire entrambe separatamente (predizione multi-output) invece di sommarle arbitrariamente. In sintesi, l'uso di questionari standard (PANAS, BDI-II, ecc.) nella predizione dell'umore fornisce una base solida e comparabile: variabili ben definite (quantitative derivanti da Likert), evidenza di validità (costrutto e criterio) e cut-off interpretativi (es. soglia clinica) che arricchiscono l'interpretazione dei modelli [46] [57].

5. Progettazione di un Survey Solido (≥170 partecipanti) e Riduzione dei Bias

La qualità dei dati raccolti tramite questionari è cruciale per costruire modelli predittivi affidabili. Per uno studio con almeno ~170 partecipanti, è fondamentale un *design del survey* accurato che massimizzi l'attenzione e la sincerità dei rispondenti, minimizzando bias come la **desiderabilità sociale**, la **cortesia verso lo sperimentatore** e le **risposte disattente o casuali**. Di seguito, le strategie chiave emerse in letteratura e nelle linee guida metodologiche:

- **Garantire anonimato e riservatezza:** L'anonimato è uno dei rimedi più efficaci contro i bias da desiderabilità sociale [58] [59]. Informare i partecipanti che il questionario è **anonimo** (nessun nome o identificativo personale legato alle risposte) e che i dati saranno trattati in forma aggregata/confidenziale li "libera" dalla preoccupazione del giudizio altrui, riducendo la tendenza a dare risposte socialmente accettabili o "di facciata". Ad esempio, Larson (2019) nota che rassicurare sulla non tracciabilità delle risposte diminuisce significativamente l'omissione di comportamenti sgradevoli riportati [59]. Nel nostro caso, assicurare che "Non esistono risposte giuste o sbagliate" e che l'obiettivo è capire il loro vero stato emotivo può incoraggiare l'onestà. Anche condurre il survey **online/autocompilato** anziché faccia-a-faccia mitiga la *courtesy bias*: il partecipante non ha di fronte un intervistatore da compiacere, quindi è più propenso a esprimere valutazioni negative o ammettere sentimenti "socialmente sconsentiti" (es. tristezza,

ansia) [60]. Se il survey fosse somministrato in presenza, è opportuno che l'ambiente sia privato e non giudicante, e che l'eventuale ricercatore adotti un atteggiamento neutro.

- **Istruzioni chiare e enfasi sull'importanza dell'attenzione:** Prima di iniziare, fornire **istruzioni dettagliate** sul completamento, sottolineando il valore di risposte accurate. Ad esempio, spiegare che "Abbiamo bisogno delle tue risposte autentiche affinché i risultati siano utili" può motivare l'impegno. Alcuni studi raccomandano di includere un **messaggio di avvertimento** sul controllo di qualità, del tipo: *"Questo questionario contiene indicatori di attenzione per assicurarci che le risposte siano serie"*. Meade e Craig (2012) hanno evidenziato che avvertire i partecipanti in anticipo della presenza di controlli riduce drasticamente l'incidenza di risposte disattente, fungendo da deterrente. Ward e Meade (2018) hanno applicato principi di psicologia sociale (impegno, identità) per prevenire il careless responding: ad esempio, far firmare virtualmente un *"impegno a rispondere con cura"* a inizio survey ha mostrato effetti positivi [61] [62]. Senza eccedere (per non spaventare il partecipante), un breve reminder iniziale del tipo *"Per favore rispondi nel modo più accurato e sincero possibile, prenditi il tempo necessario"* può aumentare l'attenzione.

- **Lunghezza e usabilità del questionario:** Un campione di ~170+ persone probabilmente verrà raccolto online; è vitale progettare il form in modo user-friendly per evitare dropout o affaticamento. Ciò include: *survey* di durata ragionevole (idealmente <15–20 minuti), suddiviso in sezioni con intestazioni chiare; interfaccia pulita (una domanda per volta su schermo per focus, oppure gruppi logici di poche domande); indicatore di progressione (% completamento) per gestire le aspettative. Domande demografiche sensibili (es. reddito) dovrebbero essere opzionali o posizionate verso la fine, per non innescare subito desiderabilità. È utile randomizzare l'ordine degli item non cruciali alla struttura, così da ridurre bias di ordine e *straightlining*. Se si somministrano più scale, intervallarle (es. non presentare 50 domande tutte con lo stesso schema di risposta in fila, ma mescolare item di scale diverse) può mantenere vivo l'interesse ed evitare risposte per pattern ripetitivo. Importante anche *testare* il survey in anticipo (pilot test) con alcune persone per identificare domande poco chiare o lungaggini.

- **Minimizzare la desiderabilità sociale nei contenuti:** Oltre all'anonimato, è possibile attenuare il bias di *social desirability* attraverso la formulazione delle domande. Strategie note includono: usare formulazioni indirette o proiettive (p.es. *"Secondo te, quante persone si sentono tristi ogni tanto senza un motivo?"* al posto di *"Ti senti mai triste senza motivo?"* – il partecipante potrebbe rispondere più onestamente in terza persona); *normalizzare* comportamenti potenzialmente stigmatizzanti nelle istruzioni (*"È normale provare certe emozioni..."*); evitare aggettivi estremi o moralmente caricati. Inoltre, si possono inserire nelle opzioni di risposta delle **scelte neutre** o equilibrate: scale con entrambe le polarità definite e includendo l'opzione "preferisco non rispondere" per domande sensibili. Un'altra tecnica è il **bogus pipeline**, utilizzato in ricerca sociale: far credere ai partecipanti che esista un meccanismo in grado di verificare la veridicità delle risposte (es. un finto "rivelatore di menzogne" a cui sono collegati) li induce paradossalmente ad essere più onesti [63]. Nel nostro contesto online ciò non è pratico, ma il concetto sottostante – *convincerli che l'onestà è attesa e che eventuali falsità sarebbero evidenti* – può essere comunicato in modo soft. Ad esempio: *"Rispondi in base a come ti senti realmente, non a come pensi che 'si dovrebbe' sentire. Ogni persona è diversa e noi apprezziamo la tua sincerità."*. Infine, si può considerare l'inclusione di una breve **scala di desiderabilità sociale** (come la Marlowe-Crowne Social Desirability Scale, versione breve ~10 item). Ciò non elimina il bias nelle risposte principali, ma permette di misurare la tendenza del partecipante a presentarsi sotto una luce positiva, e in analisi si può controllare per questo punteggio o escludere chi ottenesse un valore estremo.

- **Domande filtro e catch trials:** L'inserimento di **attenzionevoli (attention checks)** è altamente raccomandato per individuare chi compila "a caso" o senza leggere. Si possono usare *catch trials espliciti*, ad esempio: *"Questo è un controllo di attenzione: seleziona 'Molto d'accordo'."* in mezzo ad altri item Likert – chi sbaglia viene segnato. Oppure domande ovvie del tipo *"Il cielo è blu?"* con opzioni (Sì/No) – chi risponde in modo assurdo ("No" in questo caso) probabilmente non presta attenzione. Un esempio concreto: nel survey EMA di Jaso et al., valutare se $\geq$60% degli item ha la stessa risposta equivale a un *filtro ex-post* [34]; inserirne uno in situ potrebbe essere: *"Per verificare l'attenzione, scegli la terza opzione in questa scala."*. È importante posizionarli in punti non prevedibili (non tutti all'inizio o fine) e non troppo frequenti per non irritare i partecipanti (uno ogni ~15-20 item può bastare). Allo stesso modo, **domande filtro** possono servire a verificare la coerenza: ad esempio ripetere una domanda chiave in forma leggermente diversa più avanti e controllare se le risposte divergono eccessivamente (inconsistenza potrebbe segnalare disattenzione). Chi fallisce questi controlli può essere escluso dall'analisi; in uno studio con target 170 partecipanti è prudente raccoglierne qualcosa in più (es. 180-190) prevedendo che una quota (~5-10%) possa essere rimossa per inattenzione.

- **Riduzione di acquiescenza e bias di risposta:** Oltre alla desiderabilità, esistono bias come l'**acquiescenza** (tendenza a dire sempre "sì/vero") o il *contrarian bias* opposto. Per attenuare ciò, è buona norma includere item formulati in direzione opposta (*reverse-keyed items*). Ad esempio, nella scala PANAS, alternare parole positive e negative già costringe il rispondente a valutare attentamente ogni item. Nelle scale di depressione, includere item positivi (es. *"Mi sento pieno di energie"* in mezzo a enunciati negativi) rivela eventuali risposte automatiche (un partecipante poco attento potrebbe segnare alto su tutto, contraddicendosi su item invertiti). Naturalmente, gli item *reverse* devono poi essere ricodificati nel punteggio, ma aiutano a identificare protocolli inaffidabili (es.: *incoerenza even-odd* calcolata da `careless` confronta proprio item diretti vs invertiti [33]).

- **Incentivi e rapporto con i partecipanti:** Per raggiungere $\geq$170 partecipanti ed ottenere dati di qualità, può essere utile prevedere *incentivi adeguati* (crediti formativi, voucher, compenso economico simbolico) legati al completamento valido del questionario. Studi mostrano che incentivi possono aumentare l'attenzione, anche se vanno calibrati per non introdurre bias (compensi troppo alti potrebbero spingere a partecipare anche chi non è veramente interessato, generando poi risposte di scarsa qualità). Una pratica efficace è *screenare i partecipanti* in base a criteri di inclusione pertinenti: ad esempio, se interessa l'umore in popolazione generale adulta, assicurarsi che i rispondenti abbiano $\geq$18 anni e magari usare una domanda filtro iniziale ("Stai rispondendo in un luogo senza distrazioni?") per far riflettere sul contesto di compilazione. Se il reclutamento avviene online (es. piattaforme tipo Prolific, MTurk), utilizzare reputazione o punteggi di qualità dei worker come criterio per invitare (questo spesso riduce l'incidenza di careless responses).

In conclusione, un *survey design* attento unito a tecniche preventive può drasticamente migliorare la qualità dei dati: **anonimato, istruzioni chiare, lunghezza moderata, attenzione check e bilanciamento delle domande** agiscono sinergicamente per ridurre bias di desiderabilità sociale, di cortesia e inattentività. Una volta raccolti i dati, l'uso di strumenti come il pacchetto R `careless` e `EMAeval` consente di *verificare* ulteriormente la qualità (ad esempio generando report sui pattern sospetti [64]) e pulire il dataset prima dell'analisi finale. Queste accortezze garantiscono che con ~170 partecipanti si ottengano **dati validi e utilizzabili**, ponendo basi solide per implementare il workflow di analisi (ad esempio modellazione conforme e FDA) in R senza incorrere in distorsioni dovute a errori di misurazione o bias sistematici.

Bibliografia (APA)

- Dey, D., Ghosal, R., Merikangas, K., & Zipunnikov, V. (2024). *Functional principal component analysis for continuous non-Gaussian, truncated, and discrete functional data*. **Statistics in Medicine, 43**(23), 5046–5069. https://doi.org/10.1002/sim.9704 [12] [14]

- Li, Y., & Zhou, X. (2024). *Conformal Depression Prediction: A plug-and-play uncertainty quantification method for automated depression recognition*. ArXiv preprint arXiv:2405.18723 [2] [7] . *(In corso di stampa)*

- Roohi, S., Skarbez, R., & Nguyen, H. D. (2024). *Reliable uncertainty estimation in emotion recognition in conversation using conformal prediction framework*. **Natural Language Processing**, FirstView, 1–24. https://doi.org/10.1017/nlp.2024.48 [9] [10]

- Goldammer, P., Annen, H., Stöckli, P., & Jonas, K. (2020). *Careless responding in questionnaire measures: Detection, impact, and remedies*. **The Leadership Quarterly, 31**(4), 101384. https://doi.org/10.1016/j.leaqua.2019.101384 [36] [65]

- Jaso, B. A., Kraus, N. I., & Heller, A. S. (2022). *Identification of careless responding in ecological momentary assessment research: From post-hoc analyses to real-time data monitoring*. **Psychological Methods, 27**(6), 958–981. https://doi.org/10.1037/met0000312 [37] [38]

- Mor, S., Mira, A., Quero, S., García-Palacios, A., Baños, R. M., & Botella, C. (2020). *Positive and Negative Affect Schedule (PANAS): Psychometric properties of the online Spanish version in a clinical sample with emotional disorders*. **BMC Psychiatry, 20**(1), 56. https://doi.org/10.1186/s12888-020-2472-1 [47]

- Yentes, R. D., & Wilhelm, F. (2018). *careless: Procedures for computing indices of careless responding (R Package Version 1.2.2)*. Journal of Open Source Software, 3(32), 790. https://doi.org/10.21105/joss.00790 [32] [33]

Suggerimenti pratici (R workflow): Per implementare il workflow in R, si consiglia di sfruttare i pacchetti dedicati menzionati sopra. In fase di preparazione dati, utilizzare `careless` [32] per calcolare indici come longstring, IRV e Mahalanobis – es.: `careless::longstring(df)` – e rimuovere/escludere soggetti oltre soglie critiche. Per dati EMA, il pacchetto `EMAeval` (GitHub: *manateelab/EMAeval-R-Package*) offre funzioni per individuare risposte careless in serie temporali e persino generare allerte real-time [41] . Per l'analisi conformale, esistono implementazioni in R sia manuali sia tramite pacchetti: `conformal` (su GitHub [66]) implementa prediction intervals conformali oggetto-oriented, e il pacchetto `tidymodels` fornisce vignette su come integrare inference conformale nei workflow di modeling (cfr. blog *"Conformal inference for regression"* [67]). Ad esempio, dopo aver addestrato un modello `lm` o `ranger`, si può usare il pacchetto `conformalInference` (dai lavori di Lei et al.) per ottenere intervalli validati. In caso di output funzionali o curve, `conformalInference.fd` [24] permette di eseguire conformal prediction anche con risposte di natura funzionale (utile se, ad esempio, vogliamo predire l'intera traiettoria di umore di un giorno futuro). Per la parte FDA in senso stretto, utilizzare `fda.usc` per analisi esplorative (funzioni come `fdata()` per creare oggetti funzionali, `optim.basis()` per scegliere basi, `fdata.pc` per PCA funzionale). Se si vuole applicare la FPCA semiparametrica di Dey et al., gli autori hanno fornito codice (contattabile via email) e in attesa di un rilascio CRAN, si può utilizzare `mgcv` (modello additivo) con famiglia binomiale/ordinal per stimare la covarianza latente come approccio alternativo. Per cluster o visualizzazione di sequenze categoriali (es. stati d'animo sequenziali), `cfda` offre funzioni come `compute_optimal_encoding()`

e plot dei percorsi in spazi ridotti [17]. Infine, per valutare modelli e metriche: utilizzare `Metrics` o funzioni custom per calcolare PICP/MPIW su set di test predetti conformalmente, e il pacchetto `psych` per calcolare α di Cronbach e altre statistiche psicometriche sui questionari (utile per verifiche interne). Integrando questi strumenti, il workflow in R potrà: **(a)** pulire il dataset dalle risposte inattendibili, **(b)** rappresentare le risposte questionario in forma adatta (feature scalari o funzionali in base all'analisi), **(c)** addestrare modelli predittivi dell'umore con pacchetti ML (es. `caret` o `tidymodels`), **(d)** applicare conformal prediction per intervalli di confidenza affidabili, e **(e)** validare il modello (ad esempio controllando che la copertura effettiva dei prediction intervals sia conforme al livello richiesto [7]). Questo approccio end-to-end, unito alle best practice di survey design, massimizzerà le chance di ottenere predizioni dell'umore accurate e generalizzabili. [1] [64]

---

[1] [2] [3] [4] [5] [6] [7] [8] [11] [56] Conformal Depression Prediction
https://arxiv.org/html/2405.18723v3

[9] [10] Reliable uncertainty estimation in emotion recognition in conversation using conformal prediction framework | Natural Language Processing | Cambridge Core
https://www.cambridge.org/core/journals/natural-language-processing/article/reliable-uncertainty-estimation-in-emotion-recognition-in-conversation-using-conformal-prediction-framework/CE882D4B782256860B81AD2834A27477

[12] [13] [14] [21] [22] [23] [25] [26] [27] Functional Principal Component Analysis for Continuous Non-Gaussian, Truncated, and Discrete Functional Data - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC11586909/

[15] [17] Categorical Functional Data Analysis. The cfda R Package
https://www.mdpi.com/2227-7390/9/23/3074

[16] CRAN Task View: Functional Data Analysis
https://cran.r-project.org/view=FunctionalData

[18] [19] [20] Regression with Ordered Predictors via Ordinal Smoothing Splines
https://www.frontiersin.org/journals/applied-mathematics-and-statistics/articles/10.3389/fams.2017.00015/pdf

[24] CRAN: Package conformalInference.fd
https://cran.r-project.org/package=conformalInference.fd

[28] Open science perspectives on machine learning for the identification ...
https://compass.onlinelibrary.wiley.com/doi/full/10.1111/spc3.12941

[29] [PDF] Careless Responding in Survey Research: An Examination of ...
https://repository.fit.edu/cgi/viewcontent.cgi?article=1397&context=etd

[30] [32] [33] [36] careless: Procedures for Computing Indices of Careless Responding
https://cran.r-project.org/web/packages/careless/careless.pdf

[31] [40] [61] [62] [65] Dealing with Careless Responding in Survey Data: Prevention, Identification, and Recommended Best Practices | Annual Reviews
https://www.annualreviews.org/content/journals/10.1146/annurev-psych-040422-045007

[34] [35] [37] [38] [41] [64] Identification of careless responding in ecological momentary assessment research: from post-hoc analyses to real-time data monitoring - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC11565177/

[39] Identifying Careless Responding in Web-Based Surveys
https://econtent.hogrefe.com/doi/10.1027/2151-2604/a000555

42 45 46 47 Positive and Negative Affect Schedule (PANAS): psychometric properties of the online Spanish version in a clinical sample with emotional disorders - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC7008531/

43 44 PANAS Scale: The Positive & Negative Affect Schedule
https://positivepsychology.com/positive-and-negative-affect-schedule-panas/

48 Predicting Positive Psychological States using Machine Learning ...
https://www.medrxiv.org/content/10.1101/2025.04.26.25326492.full

49 50 51 53 54 55 57 Reliability and Validity of the Beck Depression Inventory-II among Korean Adolescents - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC5240453/

52 [PDF] Beck Depression Inventory-II: A Study for Meta Analytical Reliability ...
https://files.eric.ed.gov/fulltext/EJ1304998.pdf

58 Minimizing Social Desirability in Questionnaires of Non-Cognitive ...
https://www.ejper.com/minimizing-social-desirability-in-questionnaires-of-non-cognitive-measurements

59 Social-desirability bias - Wikipedia
https://en.wikipedia.org/wiki/Social-desirability_bias

60 How to Reduce Social Desirability Bias - SmartSurvey
https://www.smartsurvey.com/blog/how-to-reduce-social-desirability-bias

63 Social Desirability Bias - an overview | ScienceDirect Topics
https://www.sciencedirect.com/topics/psychology/social-desirability-bias

66 isidroc/conformal: Conformal prediction in R - GitHub
https://github.com/isidroc/conformal

67 Conformal inference for regression models - tidymodels
https://www.tidymodels.org/learn/models/conformal-regression/