QÜESTIIÓ, vol. 22, 3, p. 493-510, 1998

LA VERIFICACIÓ ALEATÒRIA: UNA ESTRATÈGIA PER MILLORAR I AVALUAR LA QUALITAT DE L'ENTRADA DE DADES

J.M. DOMÈNECH MASSONS
J.M. LOSILLA VIDAL
M. POTELL VIDAL
Universitat Autònoma de Barcelona*

S'aborda la problemàtica de la reducció dels errors que es produeixen durant la introducció de les dades i que no es poden controlar mitjançant proteccions automàtiques. Enfront aquest problema, l'estratègia habitual és la «doble entrada» (DE) de les dades, la qual augmenta considerablement el cost de la investigació. Com alternativa a aquesta estratègia es proposa un nou procediment, implementat en el Sistema DAT, que es basa en un procés de «verificació aleatòria» (VA) d'un percentatge del total de dades. A més de reduir el cost, la VA ofereix altres avantatges com el fet de proporcionar una estimació del percentatge d'errors, i oferir un índex d'aptitud i eficàcia dels operadors. A la segona part de l'article es presenten els resultats d'un experiment que recolza la hipòtesi que la VA augmenta l'eficàcia de l'entrada de dades, sense minva de l'eficiència, quan es compara amb situacions en les quals no s'aplica cap control que permeti obtenir un indicador sobre la qualitat de les dades que s'introdueixen.

Random verification: a strategy to improve and assess the quality of data entry.

Paraules clau: Qualitat de les dades; doble entrada; entrada de dades; verificació aleatòria; Sistema DAT.

Clasificación AMS: 62D99

* Dept. de Psicobiologia i de Metodologia de les Ciències de la Salut, Universitat Autònoma de Barcelona. Investigació realitzada gràcies a l'ajut DGICYT No. PM95-126 del Ministeri d'Educació i Ciència. Correspondència: J.M. Domènech. Laboratori d'Estadística Aplicada i de Modelització. Universitat Autònoma de Barcelona. Apartat de Correus 40. 08193 Bellaterra. e-mail:josep.m.domenech@uab.es.

– Rebut el març de 1997.
– Acceptat l'octubre de 1998.

INTRODUCCIÓ

En els darrers anys s'ha produït un gran avenç en les aplicacions informàtiques encaminades a facilitar el procés de dades, però es qüestiona si realment aquest tipus d'avenços reverteixen en un increment de la qualitat dels treballs que es publiquen (Cobos, 1995; Freedland i Carney, 1992; Ronel, Varley i Webb, 1993). La falta de coneixement d'alguns procediments que aquestes aplicacions automatitzen, unit a la confiança excessiva en què els resultats són correctes pel sol fet que els presenta un ordinador, poden influir negativament en la qualitat de la informació que es publica.

Les sigles GIGO («Garbage In - Garbage Out») recorden els perills que comporta aquesta situació: si s'entren dades incorrectes en una aplicació informàtica, per exemple per realitzar una anàlisi estadística, l'aplicació no deixarà d'oferir resultats encara que aquests siguin incorrectes. Una expansió més recent de GIGO és «Garbage In, Gospel Out», que és un comentari sarcàstic sobre la tendència a dipositar una confiança excessiva en les dades elaborades per l'ordinador (Dumbill, 1994).

L'«efecte GIGO» es produeix amb més freqüència en investigacions amb grans volums de dades (estudis multicèntrics, longitudinals amb molts seguiments, etc.), degut a la dificultat que representa per l'investigador controlar el correcte funcionament de les mils d'operacions que realitza amb les seves dades a través de diferents aplicacions informàtiques (Barton, Hatcher, Schurig i Marciano, 1991; Moritz *et al.*, 1995).

Ningú discuteix la importància del procés de dades cara a la validesa de les conclusions d'una investigació. No obstant això, hom acostuma a pensar només en l'anàlisi estadística, que és una de les etapes més emblemàtiques però que està molt lluny de ser la única. En aquest treball conceptualitzem el procés de dades com una activitat complexa, les principals etapes d'actuació de la qual són:

1) Disseny de la base de dades (estructura de les taules, proteccions d'entrada, etc.).
2) Entrada de les dades.
3) Preparació de la matriu de dades (generació de variables, selecció de casos, creació de taules rectangulars de dades, etc.).
4) Anàlisis de les dades.

Centrant-nos en les dues primeres etapes del procés de dades, els referents per avaluar la qualitat de les dades són: (*a*) estar exemptes d'errors; (*b*) no contenir valors omesos («missing»), i (*c*) ajustar-se als moments temporals de registre establerts en el disseny (cas d'estudis longitudinals).

Existeixen diferents aspectes que poden afectar la qualitat de les dades i, per tal d'incidir sobre ells, considerem convenient establir un primer criteri de demarcació que vindria donat per la possibilitat d'exercir un control automàtic.

Els *controls automàtics de qualitat* poden expressar la *consistència* de forma *matemàtica* donant lloc a un error, o de forma *probabilística* donant lloc a un avís. Així, un valor fora de rang (p.e. talla d'un adult igual a 7cm.), la inconsistència entre variables que mantenen relacions lògiques (p.e. sexe masculí i nombre d'embarassos diferent de «no aplicable») i el control de valors omesos, donarien lloc a *errors*, mentre que una categoria, una combinació de categories, una relació lògica o un rang de valors poc probables (p.e. adult amb talla superior a 199 cm.) donarien lloc a *avisos*.

Els *errors no controlables de manera automàtica* són aquelles unitats d'informació que compleixen les condicions de rang, lògiques i de valor «missing» anteriors, però que no coincideixen amb el valor original. Aquest tipus d'errors es poden *limitar* incorporant factors que incideixin sobre aspectes atencionals, perceptuals i motivacionals de l'operador encarregat d'introduir les dades (Henning, Sauter, Salvendy i Krieg, 1989; Pocius, 1991; González, 1993; Lalomia, i Sidowski, 1993). Les investigacions realitzades sobre aquests factors s'emmarquen dins l'àmbit d'estudi multidisciplinar de l'«Interacció Home-Ordinador» o HCI («Human Computer Interaction») (Dillon, 1983; Card, Moran i Newell, 1983; Schneiderman, 1992; Johnson, 1992; Carroll, 1993; Wallace i Anderson, 1993).

Una característica dels estudis realitzats en el context de l'HCI és que solen ser específics respecte al tipus d'aplicació informàtica i molt generals pel que fa als objectius que es desitgen aconseguir (p.e. satisfacció, comoditat, eficàcia, eficiència, etc.). En aquest àmbit la major part de treballs s'han centrat en el disseny de la interfície d'usuari, tractant aspectes com els sistemes gràfics orientats a la tasca, la disposició i el contingut dels menús, els tipus de «controls» o mecanismes que l'usuari pot utilitzar per indicar les accions que desitja realitzar, etc. (Wolf, 1992; Croquet, 1994; Howes, 1994; Karwowski, Eberts, Salvendy i Noland, 1994).

Aquest treball se centra en les aplicacions d'entrada de dades i, dins d'aquestes, en la problemàtica de la reducció dels errors d'operador no controlables de forma automàtica, mitjançant un mecanisme anomenat «verificació aleatòria» (VA). La VA forma part d'un conjunt d'aportacions que es deriven del Sistema *DAT* (Domènech i Losilla, 1995). Seguidament es presenten les característiques generals del Sistema, a continuació s'ubica la VA en relació amb altres estratègies que persegueixen objectius similars i, per últim, es descriu un estudi experimental encaminat a avaluar l'eficàcia diferencial de la VA enfront d'altres procediments.

CARACTERÍSTIQUES GENERALS DEL SISTEMA *DAT*

El Sistema *DAT* forma part d'un ampli projecte l'objectiu del qual és desenvolupar un sistema eficient pel procés de dades científiques que permeti el control (des de la seva captura fins a la seva preparació i exportació al sistema d'anàlisi estadística) de tots

els aspectes que afecten a la qualitat de les dades. La Figura 1 ubica, en el context del procés de dades, els tres mòduls que el formen (*EnDat/Lab*, *EnDat* i *ExperDat*).

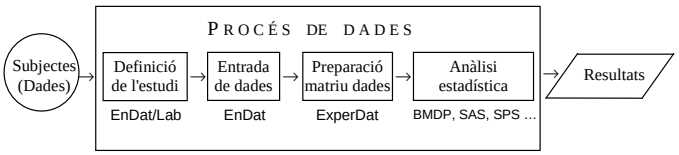

Figura 1. Etapes d'un procés de dades.

S'està preparant la versió comercial del Sistema DAT per a Microsoft Windows (3.1, 3.11, '95 i NT), a partir de les versions beta que des de fa 5 anys s'estan utilitzant de forma intensiva per a la creació, captura i gestió de bases de dades en quatre àmbits: *administració pública* (accidents laborals de Catalunya), *indústria farmacèutica* (assaigs clínics d'un important grup farmacèutic), *laboratori universitari d'estadística aplicada* (estudis per a medicina, psicologia i sociologia) i *investigació epidemiològica* (línia d'investigació sobre epidemiologia psiquiatria infantil, subvencionada amb els ajuts DGICYT PM91-209 i PM95-209).

Les principals característiques del Sistema *DAT* es poden sintetitzar exposant el seu ús en l'esmentada línia d'investigació sobre epidemiologia, que es caracteritza pel registre d'un gran nombre de variables que es corresponen amb els nombrosos ítems de cada instrument d'avaluació clínica. Per tal d'il·lustrar la complexitat i l'enorme volum de dades a manipular, així com la flexibilitat del Sistema DAT per dur a terme la seva gestió, n'hi ha prou amb dir que un d'aquests instruments diagnòstics és l'entrevista clínica estructurada DICA-R, amb més d'un miler de preguntes, que s'administra al cas d'estudi, als seus pares (per recaptar més informació sobre el subjecte) i a cadascun dels seus germans. A més, l'entrevista DICA-R es realitza en dues ocasions (test i retest) i algunes vegades es graven tant les respostes registrades per l'entrevistador com les registrades per un o varis observadors.

El mòdul *EnDat/Lab* permet especificar el disseny relacional de les dades d'aquest estudi, així com incloure totes les proteccions necessàries durant l'entrada de dades, els salts que guien l'entrevistador sobre les preguntes a efectuar, i també definir els camps virtuals amb els diagnòstics DSM-IV, parcials i finals, necessaris per administrar correctament l'entrevista. L'*EnDat/Lab* genera un conjunt d'arxius que contenen aquestes especificacions, i que seran llegits pel mòdul *EnDat*, encarregat de proporcionar la interfície d'usuari per dur a terme l'entrada de dades. Per últim, el mòdul *ExperDat* permet generar noves variables, definir fàcilment *vistes* que reorganitzin l'estructura relacional de la base de dades i preparin matrius rectangulars, les

quals poden implicar fins i tot una nova definició del *cas d'estudi* per adaptar-la al *cas d'anàlisi* d'un determinant objectiu d'investigació. Així, les entrevistes de cada germà, per exemple, a cops es poden considerar rèpliques de l'entrevista del subjecte, mentre que altres vegades constitueixen entrevistes de diferents subjectes (de la Osa, Ezpleleta, Doménech, Navarro i Losilla, 1996).

El Sistema *DAT* és també una plataforma idònia per la recerca en l'àmbit de la HCI, especialment pel que fa a: (*a*) detecció dels aspectes que causen error i/o redueixen la velocitat del treball dels operadors; (*b*) avaluació de noves alternatives que evitin els errors i incrementin l'eficiència, entre les quals es troba la VA, i (*c*) desenvolupament d'instruments indicatius de la destresa dels operadors, que poden facilitar la fase de selecció de personal.

LA VERIFICACIÓ ALEATÒRIA COM ALTERNATIVA O COMPLEMENT A LA DOBLE ENTRADA

Com ja dèiem a la introducció, les implicacions de l'«efecte GIGO» s'aguditzen quan els estudis són d'una certa magnitud. En aquests casos és imprescindible disposar d'un indicador de la qualitat de les dades, el qual és impossible d'obtenir si es realitza únicament una «entrada simple» (SE). En front aquest problema, l'alternativa que es planteja habitualment és la «doble entrada» (DE) de les dades, que implica que s'introdueixi dues vegades cada registre. Aquesta estratègia es concep com l'única que assegura que les dades estiguin lliures d'errors abans de procedir a la seva anàlisi, al mateix temps que permet obtenir un índex de la qualitat del treball dels operadors (Blumenstein, 1993; Gassman, Owen, Kuntz, Martin i Amoroso, 1995; Gibson, Harvey, Everett i Parmar, 1995; Wolf, 1993). S'ha de tenir en compte que si es desitja aprofitar aquest avantatge abans de finalitzar la recollida de dades (cas especialment important en els estudis longitudinals), o si els qüestionaris en paper només poden estar en el lloc on s'introdueixen les dades durant un temps limitat (com succeeix, per exemple, quan existeixen motius de seguretat), la DE no es pot realitzar en diferit, sinó que s'ha d'anar aplicant durant el procés d'entrada de dades (al finalitzar un bloc de registres, al finalitzar el dia, per un operador que fa les funcions de supervisor del què realitza la primera entrada, etc.).

Per tant, el principal desavantatge que comporta l'aplicació de la DE és, sens dubte, el seu elevat cost, tant econòmic com de temps. Per aquest motiu, en la pràctica, només els estudis que compten amb recursos econòmics suficients (p.e. els assaigs clínics) poden aplicar la DE, quedant la majoria d'estudis desproveïts de mecanismes que assegurin la qualitat de les dades, i exposats per tant a l'efecte GIGO.

Enfront d'aquesta situació, Domènech i Losilla (1995) proposen una estratègia de «verificació aleatòria» (VA) de les dades, que consisteix en la doble entrada d'un

percentatge del total de dades, elegides a l'atzar pel sistema informàtic i sol·licitades amb un *retard* suficient per tal d'evitar efectes de record, i en la reproducció dels errors comesos per l'operador durant la primera entrada.

El principal avantatge de la VA és que redueix considerablement el cost de l'entrada de dades mantenint gran part de les característiques positives que s'atribueixen a la DE realitzada en condicions òptimes. És a dir, la VA permet obtenir, en qualsevol moment del procés d'entrada de dades, una estimació del *percentatge d'errors* de les dades i un *índex de l'aptitud i de l'eficàcia dels operadors*.

Un aspecte a destacar aquí és que la VA és compatible amb la DE, perquè es pot utilitzar la VA durant la fase de recollida de dades, i al finalitzar aquesta fase completar la DE introduint els registres restants. Tanmateix, la VA implica necessàriament que sigui el mateix operador qui realitzi la reeintroducció de les dades; no obstant això, té l'avantatge de proporcionar a l'operador un «feed-back» dels errors que es van produint.

També és possible combinar la DE amb la VA en estudis amb estrats de subjectes que per les seves característiques requereixen diferents nivells de control. Per exemple, en l'anomenat registre anual d'accidents laborals, convé aplicar DE o VA del 100% a tots els comunicats d'accidents mortals i greus (aprox. 2.000) i utilitzar un percentatge de VA molt més baix pels accidents lleus (aprox. 160.000), reduint d'aquesta forma el cost global del registre, proporcionant un índex de la seva qualitat i contribuint a aconseguir l'estàndard de qualitat establert.

Aquest «feed-back» de la VA ens permet plantejar la primera hipòtesi del nostre estudi:

Hipòtesi 1: *El «feed-back», inherent a la situació de VA, pot contribuir positivament a incrementar l'eficàcia global del procés d'introducció de dades, és a dir, el nombre d'errors comesos serà menor en les situacions d'entrada de dades en les quals l'operador rep «feed-back».*

Des d'un punt de vista pràctic i atenent només al cost econòmic de l'entrada de dades, també té interès comparar l'*eficiència* dels operadors en situacions en què no rebin «feed-back» sobre els errors que cometen, amb la seva eficiència en situacions en què sí el reben. Però la comparació dels nivells d'eficiència implica contrastar prèviament la següent hipòtesi:

Hipòtesi 2: *El temps que s'inverteix en la introducció de qüestionaris de dades sense errors no és superior al que s'inverteix en la introducció de qüestionaris de dades amb errors, independentment de si la situació comporta o no «feed-back» per a l'operador.*

En efecte, una diferència en el sentit contrari a l'expressat per aquesta hipòtesi, és a dir, major temps requerit per introduir dades sense errors que dades amb errors, implicaria que un augment en l'eficàcia comporta necessàriament una disminució en l'eficiència i, en conseqüència, que no tingui cap valor comparar les eficiències dels operadors en les diferents situacions d'entrada de dades, sinó únicament els seus nivells d'eficàcia. Si, pel contrari, les dades no contradiuen la hipòtesi 2, llavors contrastarem la següent hipòtesi sobre l'eficiència:

Hipòtesi 3: *El «feed-back» associat a la VA no comporta una minora en l'eficiència, és a dir, el nombre de dades per unitat de temps introduïdes sense cometre errors és igual o major en les situacions en les quals l'operador rep «feed-back» que en les que no en rep.*

MÈTODE

Subjectes

Els subjectes són 45 estudiants voluntaris de primer curs de Psicologia de la Universitat Autònoma de Barcelona, amb edats compreses entre els 18 i els 20 anys i un 67% de dones, que s'han dividit a l'atzar en tres grups de mida 15.

Als subjectes se'ls ha ofert, com a recompensa per la seva participació, una llicència d'ús d'un programari informàtic comercial valorat en deu mil pessetes, així com l'accés a l'aula de microinformàtica del nostre laboratori durant el proper curs acadèmic.

S'ha considerat com a condició d'admissió que el subjecte tingui un nivell suficient d'habilitat en l'ús d'un ordinador personal (equivalent al necessari per escriure un document simple amb un programari de tractament de textos).

Material

S'han omplert a mà 100 qüestionaris amb dades fictícies i, per gravar-los, s'ha construït amb el Sistema *DAT* un formulari informatitzat sense proteccions per rangs, condicions lògiques, etc., la qual cosa permet introduir qualsevol valor. Per incrementar la taxa d'errors, s'ha dissenyat la pantalla de manera que la disposició dels camps de captura no coincideixi amb els qüestionaris impresos. L'experiment s'ha realitzant en el nostre laboratori, en grups de 15 subjectes, utilitzant ordinadors PC-486/66, monitor color de 14" i tarja gràfica VGA.

Disseny

S'ha aplicat un disseny de grups a l'atzar amb registre pretest. La variable independent és el tipus de situació d'entrada de dades assignada a cada grup: «verificació aleatòria» (VA), «placebo» (PL) i «control» (CO).

La *condició VA* es caracteritza per una consigna verbal informant que el programari controlarà els errors comesos durant l'entrada de dades i obligarà a repetir la introducció d'alguns formularis. El percentatge de formularis a verificar es fixa en un 15% i el retard en la doble entrada d'aquest percentatge de casos s'estableix en dos formularis.

La condició *PL* es caracteritza només per una consigna verbal informant que durant l'entrada de dades el programari controlarà el nombre d'errors comesos. Aquesta condició és necessària per distingir els efectes específics del «feed-back» que comporta la VA de l'efecte degut a la coneixença que s'estan registrant els errors comesos.

En la *condició CO* els subjectes no reben cap consigna.

Hem definit quatre variables dependents: *mesura de l'eficàcia* (proporció de camps de dades introduïts sense error), *mesura de l'eficiència* (nombre de registres de dades o de qüestionaris per hora de treball introduïts sense error), *temps mitjà per introduir els qüestionaris amb error* (en segons) i *temps mitjà per introduir els qüestionaris sense error* (en segons). El temps de la segona entrada dels qüestionaris revisats mitjançant VA no ha estat considerat perquè aquest temps no existeix en les altres dues condicions.

Procediment

Cada grup de 15 subjectes ha introduït a l'ordinador les dades del conjunt de qüestionaris, en una sessió que ha durat aproximadament dues hores. En la fase de reclutament la tasca a realitzar no es presenta com un experiment sinó com una col·laboració en un treball d'investigació real. Un cop finalitzada la recollida de dades s'ha comunicat als participants els objectius reals de l'estudi (atenent a l'article 34/89 del codi de deontologia del Col·legi de Psicòlegs de Catalunya).

La sessió s'ha estructurat en tres fases: entrenament, línia base i experimental. La *fase d'entrenament* té com a objectius: (1) informar els subjectes de quina serà l'operativa a seguir per entrar els qüestionaris a l'ordinador, (2) anticipar que està previst realitzar parades durant la sessió; (3) incentivar la rapidesa de l'entrada de dades, oferint un premi addicional als subjectes que introdueixin un major nombre de qüestionaris durant la sessió; i (4) que el subjecte practiqui amb el programari informàtic utilitzat

per l'entrada de dades. Les dades obtingudes en aquesta fase no s'han incorporat en les anàlisis dels resultats de l'estudi.

Passats els 15 minuts que ha durat la fase d'entrenament, es realitza una breu parada i, a continuació, s'inicia la *fase de línia base*, que té una durada de 45 minuts i en la qual no es realitza cap mena d'indicació als subjectes.

Finalitzada aquesta fase, s'inicia un nou descans en el qual es donen les consignes verbals corresponents a cadascuna de les tres condicions. La fase experimental té una durada de 60 minuts per als grups CO i PL, i de 75 minuts pel grup VA.

RESULTATS

Per tal de contrastar la primera hipòtesi, relativa a l'efecte del «feed-back» associat a la situació de VA sobre l'eficàcia global de l'operador (EFICACIA), s'ha ajustat un model de regressió considerant com a variable independent la situació d'entrada de dades (GRUP) descomposta en dues variables fictícies codificades respecte a la categoria de referència VA, i incloent com a variable de control el nombre de camps de dades introduïdes sense error durant la línia base (LBEFICA) juntament amb els corresponents termes d'interacció perquè no es pot descartar a priori la seva existència:

$$\begin{aligned} \text{EFICACIA} &= \alpha + \beta_1 \times \text{GRUP}_{\text{CO}} + \beta_2 \times \text{GRUP}_{\text{PL}} + \gamma \times \text{LBEFICA} + \\ &+ \delta_1 \times \text{GRUP}_{\text{CO}} \times \text{LBEFICA} + \delta_2 \times \text{GRUP}_{\text{PL}} \times \text{LBEFICA} + \varepsilon \end{aligned}$$

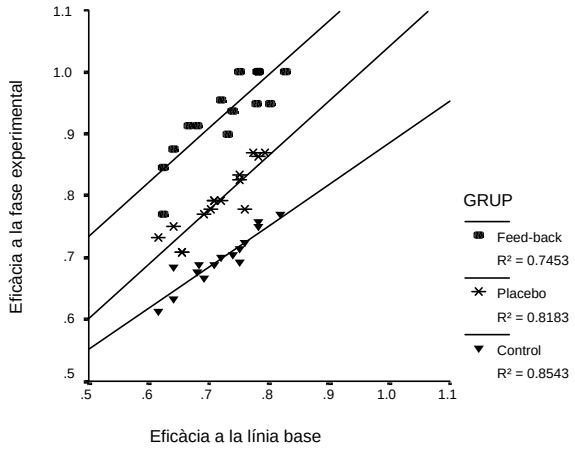

Figura 2. Relació entre l'eficàcia durant l'entrada de dades en la fase de línia base i en la fase experimental.

Els resultats de l'anàlisi (Figura 2) posen de manifest que la interacció entre GRUP i LBEFICA no és estadísticament significativa ($F(2,\ 39) = 1.03;\ p = 0.37$), motiu pel qual hem estimat un nou model sense interacció. En aquest nou model el terme d'ajust (línia base) produeix canvis molt poc importants en els coeficients *b*, però s'ha retingut perquè millora la precisió de les seves estimacions (ambdós errors estàndard disminueixen de 0.02 a 0.01).

Els resultats de l'anàlisi indiquen un *increment d'eficàcia* perquè en el grup VA s'han introduït un 27.7% (IC95%: 20.7 a 24.7) més de camps de dades sense error que en el grup CO, i un 13.1% (IC95%: 11.1 a 15.0) més de dades sense error que en el grup PL. Aquests resultats són congruents amb la hipòtesi plantejada, ja que posen de manifest l'efecte positiu sobre l'eficàcia global del procés d'introducció de dades degut al «feed-back» inherent a la situació de VA, distingint aquest efecte de l'atribuïble al fet que els subjectes puguin tenir coneixement (grup PL) que s'estan registrant els seus errors.

Per tal de contrastar la segona hipòtesi, referida a la igualtat entre els temps requerits per a la introducció de les dades sense errors respecte a les dades amb errors, s'han comparat els temps mitjans amb una prova t per a mostres relacionades. Els resultats de l'anàlisi indiquen que si bé el temps mitjà d'introducció de dades amb error en el grup VA ha incrementat significativament en 0.5 seg. ($p = 0.02$), aquest augment manca d'importància pràctica (IC95%: 0.1 a 0.9). Per tant, aquest resultat justifica el plantejament de la tercera hipòtesi.

Per tal de contrastar la tercera hipòtesi sobre l'efecte del «feed-back» inherent a la situació de VA sobre l'eficiència, s'ha procedit a estimar l'anterior model de regressió utilitzant en aquest cas com a variable dependent la mesura de l'eficiència.

Els resultats de l'anàlisi posen de manifest que la interacció entre el grup i la eficiència durant la línia base no és estadísticament significativa ($F(2,\ 39) = 0.63;\ p = 0.54$). En el model final s'ha inclòs el terme control (línia base) perquè, malgrat no es produeixen canvis apreciables en els coeficients de regressió que mesuren els efectes, millora la seva precisió. Els resultats d'aquesta última anàlisi són congruents amb la hipòtesi de manteniment de l'eficiència perquè el temps mitjà d'introducció de dades sense error, encara que és lleugerament superior en el grup VA respecte al del grup CO, comporta una diferència que no és pràcticament important (d=0.8 seg.; IC95%: (-0.5 a 2.1), ni estadísticament significativa (p=0.22). El mateix passa en el grup PL (d=0.3 seg.; IC95%: (-1.0 a 1.6; p=0.64).

DISCUSSIÓ

Els resultats trobats indiquen que la VA augmenta l'eficàcia del procés d'entrada de dades, sense minvar la seva eficiència, quant es compara amb situacions en les quals

no s'aplica cap control que permeti obtenir un indicador sobre la qualitat de les dades que s'introdueixen. A més, els resultats obtinguts al comparar el grup de VA amb el grup placebo, recolzen la hipòtesis plantejada sobre l'efecte específic del «feed-back» inherent a la situació de VA com a mecanisme explicatiu de la reducció de l'error.

Encara que l'estratègia de VA, com succeeix amb la DE, té un cost de temps més elevat que la SE, les dades aportades en el present estudi suggereixen que aquest increment del cost no és motiu suficient per excloure controls d'aquest tipus durant la introducció de les dades. Segons la nostra opinió, i per evitar en part l'«efecte GIGO», mai s'hauria d'analitzar unes dades de les quals no es disposi, com a mínim, de l'estimació de l'error que contenen, motiu pel qual considerem de gran interès pràctic els resultats trobats sobre l'eficàcia i l'eficiència de l'aplicació de controls mitjançant VA.

Per últim, cal destacar també que el Sistema *DAT* es mostra com una plataforma molt útil per a implantar models de laboratori vàlids, adreçats a avaluar l'impacte d'altres mètodes de reducció dels errors comesos pels operadors durant l'entrada de dades.

REFERÈNCIES

[1] **Barton, C., Hatcher, C., Schurig, K.** i **Marciano, P.** (1991). «Managing data entry of a large-scale interview project with optical scanning hardware and software». 20th Annual Meeting of the Society for Computers in Psychology (1990, New Orleans, Louisiana). *Behavior Research Methods, Instruments, and Computers*, **23 (2)**, 214-218.

[2] **Blumenstein, B.A.** (1993). «Verifying keyed medical research data». *Statistics in Medicine*, **12**, 1535-1542.

[3] **Card, S.K., Moran, T.P.** i **Newell, A.** (1983). *The psychology of human computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

[4] **Carroll, J.M.** (1993). «Creating a design science of human-computer interaction». *Interacting with Computers*, **5(1)**, 3-12.

[5] **Cobos, A.** (1995). «El síndrome "GIGO"». *JANO*, **49(1123)**, 481-482.

[6] **Croquet, M.** (1994). *PC y Robótica. Técnicas de interfaz*. Madrid: RA-MA.

[7] **Col·legi de Psicòlegs de Catalunya.** (1989). *Código deontológico*.

[8] **de la Osa, N., Ezpeleta, L., Doménech, J.M., Navarro, J.B.** i **Losilla, J.M.** (1996). «Fiabilidad entre entrevistadores de la DICA-R». *Psicothema*, **6(2)**, 359-368.

[9] **Dillon, R.F.** (1983). «Human factors in user-computer interaction: An introduction». *Behavior Research Methods, Instruments and Computers*, **15**, 195-199.

[10] **Domènech, J.M.** i **Losilla, J.M.** (1995). *Sistema DAT: gestor de datos científicos. Manual de referencia*. Campus de Bellaterra, Barcelona: Laboratori d'Estadística Aplicada i de Modelització. Universitat Autònoma de Barcelona.

[11] **Dumbill, E.** (1994). *Jargon Lexicon Main Index*. Diccionari en format hipertexte que es pot obtenir a Internet en la següent direcció: http://www-i5.informatik.rwth-aachen.de/mbp/jargon300/main.html.

[12] **Gassman, J.J, Owen, W.W., Kuntz, T.E, Martin, J.P.** i **Amoroso, W.P.** (1995). «Data quality assurance, monitoring, and reporting». *Controlled Clinical Trials*, **16 (2 Suppl.)**, 104S-136S.

[13] **Gibson, D., Harvey, A.J., Everett, V.** i **Parmar, M.K.** (1995). «Is double data entry necessary? The CHART trials. CHART Steering Committee. Continuous, Hyperfractionated, Accelerated Raiotherapy». *Controlled Clinical Trials*, **15 (6)**, 482-488.

[14] **González, H.** (1993). «Psicología de las interfaces: usuario-sistema: teorías y métodos». *Revista Intercontinental de Psicología y Educación*, **6(1-2)**, 35-61.

[15] **Freedland, K.E.** i **Carney, R.M.** (1992). «Data management and accountability in behavioral and biomedical research». *American Psychologist*, **47(5)**, 640-645.

[16] **Henning, R.A., Sauter, S.L., Salvendy, G.** i **Krieg, E.F.Jr.** (1989). «Microbreak length, performance, and stress in a data entry task». *Ergonomics*, **32**, 855-864.

[17] **Howes, A.** (1994). *A model of the adquisition of menu knowledge by examination*. En B. Adelson, S. Dumais y J. Olson (Eds.), *CHI'94*, pp. 445-451. Boston, MA: ACM.

[18] **Johnson, P.** (1992). *Human Computer Interaction: Psychology, Task Analysis and Software Engineering*. Londres: McGraw-Hill.

[19] **Karwowski, W., Eberts, R, Salvendy, G.** i **Noland, S.** (1994). «The effects of computer interface design on human postural dynamics: Special Issue: Festschrift in honour of Professor E. Negel Corlett». *Ergonomics*, **37(4)**, 703-724.

[20] **Lalomia, M.J.** i **Sidowski, J.B.** (1993). «Measurements of computer anxiety: A review». *International Journal of Human Computer Interaction*, **5(3)**, 239-266.

[21] **Moritz, T.E., Ellis, N.K., Villanueva, C.B., Steeger, J.E., Ludwig, S.T.** i **Deegan, N.I.** (1995). «Development of an interactive data base management system for capturing large volumes of data». *Medical Care*, **33** (10 Suppl.).

[22] **Pocius, K.E.** (1991). «Personality factors in human-computer interaction: A review of the literature». *Computers in Human Behavior*, **7(3)**, 103-135.

[23] **Ronel, R.K.; Varley, S.A.** i **Webb, C.F.** (1993). *Clinical data management*. Chichester: John Wiley & Sons.

[24] **Schneiderman, B.** (1992). *Designing the user interface: Strategies for effective human-computer interaction* (2ª ed.). Reading, MA: Addison-Wesley.

[25] **Wallace, M.D.** i **Anderson, T.J.** (1993). «Approaches to interface design». *Interacting with Computers*, **5(3)**, 259-278.

[26] **Wolf, C.G.** (1992). «A comparative study of gestual, keyboard, and mouse interfaces». *Behaviour and Information Technology*, **11(1)**, 13-23.

[27] **Wolf, R.M.** (1993). «Data quality and norms in international studies. Special Issue: Mandatory testing: Issues in policy-driven assessment». *Measurement and Evaluation in Counseling and Development*, **26 (1)**, 35-40.

ENGLISH SUMMARY

RANDOM VERIFICATION: A STRATEGY TO IMPROVE AND ASSESS THE QUALITY OF DATA ENTRY

J.M. DOMÈNECH MASSONS
J.M. LOSILLA VIDAL
M. POTELL VIDAL
Universitat Autònoma de Barcelona*

The paper addresses the problem of the reduction of errors produced during data entry which are not controllable through automatic protections. The habitual strategy in front of this problem is the «double entry» (DE) of data, which considerably increases the research cost. As an alternative to this strategy we propose a new procedure, implemented in DAT System, based on a process of «random verification» (RV) or a global data percentage. Besides of cost reduction, RV offers other advantages, as the fact of providing an error percentage estimation which supports the hypothesis that RV increases the effectiveness of data entry, without decreasing the efficiency, when is compared with situations in which there is no control about quality of entered data.

Keywords: Data quality, double entry, data entry, random verification, DAT System.

AMS Classification: 62D99

* Dept. de Psicobiologia i de Metodologia de las Ciències de la Salut, Universitat Autònoma de Barcelona. Aquesta investigació s'ha realitzat gràcies a l'ajut DGICYT No. PM95-126 del Ministeri d'Educació i Ciència.

– Received March 1997.
– Accepted October 1998.

INTRODUCTION

This work studies data entry applications and, concretely, the problematic of reduction of operators errors that are not automatically controllable, through a mechanism called «random verification» (RV). RV is one of a contribution set derived from DAT System (Domènech & Losilla, 1995), which are presented as a complement to the double data entry (DE) strategy.

Errors not automatically controllable are those units of information that fulfill previous rank, logical and missing value conditions, but that are not coincident with original value. This kind of errors could be limited by incorporating factors that affect on attentional, perceptional and motivational aspects of the operator managing data entry (Henning, Sauter, Salvendy and Krieg, 1989; Pocius, 1991; González, 1993; Lalomia and Sidowski, 1993).

GENERAL CHARACTERISTICS OF DAT SYSTEM

DAT System is an element of a wide project which has as objective the development of an efficient system for scientific data managing, allowing to control (from its capture to its preparation and exportation to the statistical analysis system) all the questions affecting data quality. Figure 1 locates, in data managing process, the 3 modules of DAT System (*EnDat/Lab, EnDat and ExperDat*).

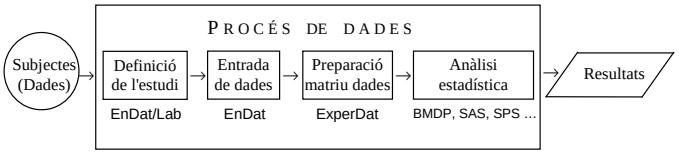

Figure 1

DAT System is also a suitable platform for research in Human Computer Interaction, specially regarding to: (*a*) detection of aspects that cause errors and/or decrease operators' work speed; (*b*) assessment of new alternatives in order to avoid errors and increase efficiency, as RV is, and (*c)*) development of tools indicating operators' kills, which could facilitate workers selection phase.

RANDOM VERIFICATION AS AN ALTERNATIVE OF COMPLEMENT TO DOUBLE ENTRY

The main disadvantage of DE is, without a doubt, its high economic and temporal cost. For this reason, in practice, only studies with huge economic resources (i.e. clinical trials) could apply double entry, being the greater of studies devoid of mechanisms that guarantee data quality.

In front of this situation, Domènech and Losilla (1995) propose a strategy of «random verification» (RV) of data, consisting on double entry of only a percentage of the total data, selected at random by computer system and requested with enough delay as to avoid memory effects, and on the reproduction of the operators' errors committed during first entry.

The main advantage of RV is that considerably decreases data entry cost, preserving a great part of positive characteristics attributed to double entry when is carried out in optimum conditions. That is to say, RV has the advantage of giving to the operator a feedback of errors produced and allows to obtain, in any time of data entry process, an estimation of data errors percentage and an index of operator' aptitude and efficacy.

This feedback of RV allow us to pose the first hypothesis of our study:

Hypothesis 1: *The feedback, inherent to RV situation, could positively contribute to increase the global efficacy of data entry process, that is to say, the number of errors committed will be less in data entry situations in which operator receives feedback.*

From a practical point of view and attending only to the economic cost of data entry, it is also interesting to compare operator' efficiency in situations in which they do not receive feedback about committed errors, with their efficiency in situations in which they do. But the comparison of efficiency levels involves the next hypothesis:

Hypothesis 2: *The time spent on the introduction of data questionnaires without errors will not be superior to the time spent on the introduction of data questionnaires with errors, independently of if the situations implicates or not feedback to the operator.*

Indeed, a difference in the opposed direction to the expressed in the hypothesis, that is to say, a greater time needed to introduce data without errors than data with errors, would implicate that increasing efficacy implies necessarily decreasing efficiency and, consequently, that comparing operators' efficiency in different data entry situations would not have any worth, but only comparing their efficacy levels. Contrarily, if data do not contradict hypothesis 2, we will contrast the next hypothesis about the efficiency:

Hypothesis 3: *The feedback associated to RV does not imply a decrease in efficiency, that is to say, the number of data per time unit introduced without error is equal or greater in situations in which operator receives feedback than in situations in which do not.*

METHOD

Subjects

The subjects are 45 voluntary students of first course of psychology at the Universitat Autònoma de Barcelona, with ages between 18 and 20 years-old and a 67% of woman, which have been divided in three groups of size 15.

Material

100 questionnaires have been hand filled with fictitious data and, to register them, a computerised form without protections has been created with DAT System.

Design

A random groups with pre-test registration design has been applied. The independent variable is the type of data entry situation assigned to each group: «random verification» (RV), «placebo» (PL) and «control» (CO).

RV condition is a verbal instruction informing that software will control the errors committed during data entry, and will oblige to repeat the introduction of some forms. The percentage of forms to verify is set at 15% and the delay in the double entry of these forms is established in two forms.

PL condition is a verbal instruction informing that software will control the number of committed errors. This condition is necessary to distinguish the specific effects of feedback that RV implies from the effect due to the knowledge that committed errors are being registered.

In Co conditions the subjects do not receive any instruction.

We have defined four dependent variables: *efficacy measurement* (proportion of data fields introduced without error), *efficiency measurement* (number of data records or questionnaires per hour of work introduced without error), *average time employed to introduce the questionnaires with error* (in seconds) *and average time employed to introduce the questionnaires without error* (in seconds). The time employed in the

second entry or revised questionnaires in RV has not been considered because this time does not exist in the other two conditions.

Procedure

Each group of 15 subjects has introduced to the computer the data from the whole set of questionnaires, in a structured session with three phases: training, baseline and experimental.

RESULTS

The results of the analysis show an efficacy increment because in RV group a 27.7% (IC95%: 20.7 to 24.7) more of data fields without error than in CO group have been introduced, and a 13.1% (IC95%: 11.1 to 15.0) more of data without error than in PL group.

The results also show that, although the average time of introduction of data with error in RV group has significantly increased in 0.5 seconds ($p = 0.02$), this increment has not realistic importance (IC95%: 0.1 to 0.9).

The results of the last analysis are congruent with the hypothesis of efficiency maintenance because the average time of introduction of data without error, although is slightly greater in RV group than in CO group, implies a difference neither of realistic importance ($d = 0.8$ sec.; IC95%: -0.5 to 2.1) nor statistically significant ($p = 0.22$). The same occurs in PL group ($d = 0.3$ sec.; IC95%: -1.0 to 1.6; $p = 0.64$).

DISCUSSION

The results show that RV increases the efficacy of data entry process, without decreasing its efficiency, when is compared with situations in which there is no control about data quality. Furthermore, the results obtained comparing RV and PL groups support the hypothesis about the specific effect of feedback inherent to RV situation as an explanatory mechanism of error reduction.

Although RV strategy, as happen to DE, has a higher time cost than simple entry, the contributions of this study suggest that this increment is not a sufficient motive to exclude this type of controls during data entry. In our opinion, it should never be analyzed data without having, at least, an estimation of the errors that they contain. For this reason, we consider the results about efficacy and efficiency in the application of controls through RV of great realistic interest.